

(c) 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works."

One-shot Learning of Human Activity with a MAP Adapted GMM and Simplex-HMM

Mario Rodriguez, Carlos Orrite, *Member, IEEE*, Carlos Medrano, *Member, IEEE*,
Dimitrios Makris, *Member, IEEE*,

Abstract—This paper presents a novel activity class representation using a single sequence for training. The contribution of this representation lays on the ability to train an one-shot learning recognition system, useful in new scenarios where capturing and labelling sequences is expensive or impractical. The method uses a Universal Background Model of local descriptors obtained from source databases available on-line and adapts it to a new sequence in the target scenario through a Maximum a Posteriori adaptation. Each activity sample is encoded in a sequence of normalized Bag of Features and modelled by a new Hidden Markov Model formulation, where the EM algorithm for training is modified to deal with observations consisting in vectors in a unit simplex. Extensive experiments in recognition have been performed using one-shot learning over the public datasets Weizmann, KTH and IXMAS. These experiments demonstrate the discriminative properties of the representation and the validity of application in recognition systems, achieving state-of-the-art results.

Index Terms—Activity Recognition, Soft-assignment, HMM, MAP Adaptation, Transfer Learning.

I. INTRODUCTION

MACHINE Learning advances in video-based human activity recognition and the ubiquity presence of video cameras in our daily life have inspired multiple application areas such as search engines, indexing, surveillance, entertainment and home automation. However, further research in action recognition is required to achieve commercially acceptable reliability.

Most of the proposed recognition approaches are trained with large amount of labelled examples using large databases and usually validating the method with leave-one-out or train-test split strategies. In practice, this is reasonable for applications with unconstrained scenarios such as searching specific activities in movies, or indexing Internet videos, where the training examples can be obtained relatively easy from on-line videos. Results in large and unconstrained datasets such as HMDB51 [1] or OlympicSports [2] are useful for general evaluation, because their examples have been collected from diverse sources, for instance Youtube or extracted from

movies, but accuracy of algorithms is not yet at the level required in many commercial applications.

Higher accuracy can be achieved in constrained scenarios and with fixed cameras, such as the ones represented by the Weizmann [3], KTH [4] or IXMAS [5] datasets, assuming the availability of several labelled examples for training.

Recognizing in a fixed scenario has the advantage of suppressing in some degree the clutter introduced by the change of background and viewpoint and therefore higher reliability may be achieved. Many applications, such as visual monitoring of elderly or disabled people at home, video surveillance or gaming, can be considered that are similarly constrained, assuming that they were installed in a specific environment. However, after the installation, the system should be re-trained again as any previously collected sequences may not be representative of the new environment. Although the performance is constrained by the number of labelled sequences used for training, collecting and labelling large amount of data for the particular scenario is infeasible, as it is laborious and may require the involvement of the user. Little research has been done in training an activity recognition system with limited number of labelled examples although being an essential feature in many practical situations [6] [7] [8] [9].

In the ideal case, only one sequence per class should be enough for activity representation as shown in some previous work [6]. However, it is important to mention that the description of an one-shot learning approach differs among papers in the literature. In order to have a better understanding of the meaning we classify the different approaches into two groups. First, the **Strict One-Shot learning** assumes only one training example available which is used to model a single class. After training several models (one per available example) of different classes separately, it is possible to combine these models in order to train a recognition systems. *Seo and Milanfar* [6] proposed a nearest-neighbour classification using a Strict One-Shot learning approach. Second, the **Relaxed One-Shot learning** process uses simultaneously multiple training examples available, assuming one per class. This relaxation allows sharing some information among the examples in order to model the classes or directly training a recognition system. Methods by *Yang et al.* [7] and *Orrite et al.* [8] follow this description as they create a vocabulary of features using sequences of the different classes. The Relaxed approach usually gives better results but at the expense of retraining the system with each new inclusion and with the inconvenience of requiring several examples from the beginning. In *Rodriguez*

Mario Rodriguez and Carlos Orrite are with CVLab, I3A, Zaragoza University, c/Mariano Esquillor s/n 50018, Zaragoza, Spain.

Carlos Medrano is with EduQTech, E.U. Politecnica, Zaragoza University, c/Ciudad Escolar s/n, 44003, Teruel, Spain.

Dimitrios Makris is with Digital Imaging Research Centre, Kingston University, Penrhyn Road Kingston upon Thames Surrey KT1 2EE, UK.

This work was partially supported by Spanish Grant TIN2013-45312-R (MINECO), FEDER, Gobierno de Aragon and the European Social Found. Mario Rodriguez was sponsored by Spanish FPI Grant BES-2011-043752 and EEBB-I-14-08410.

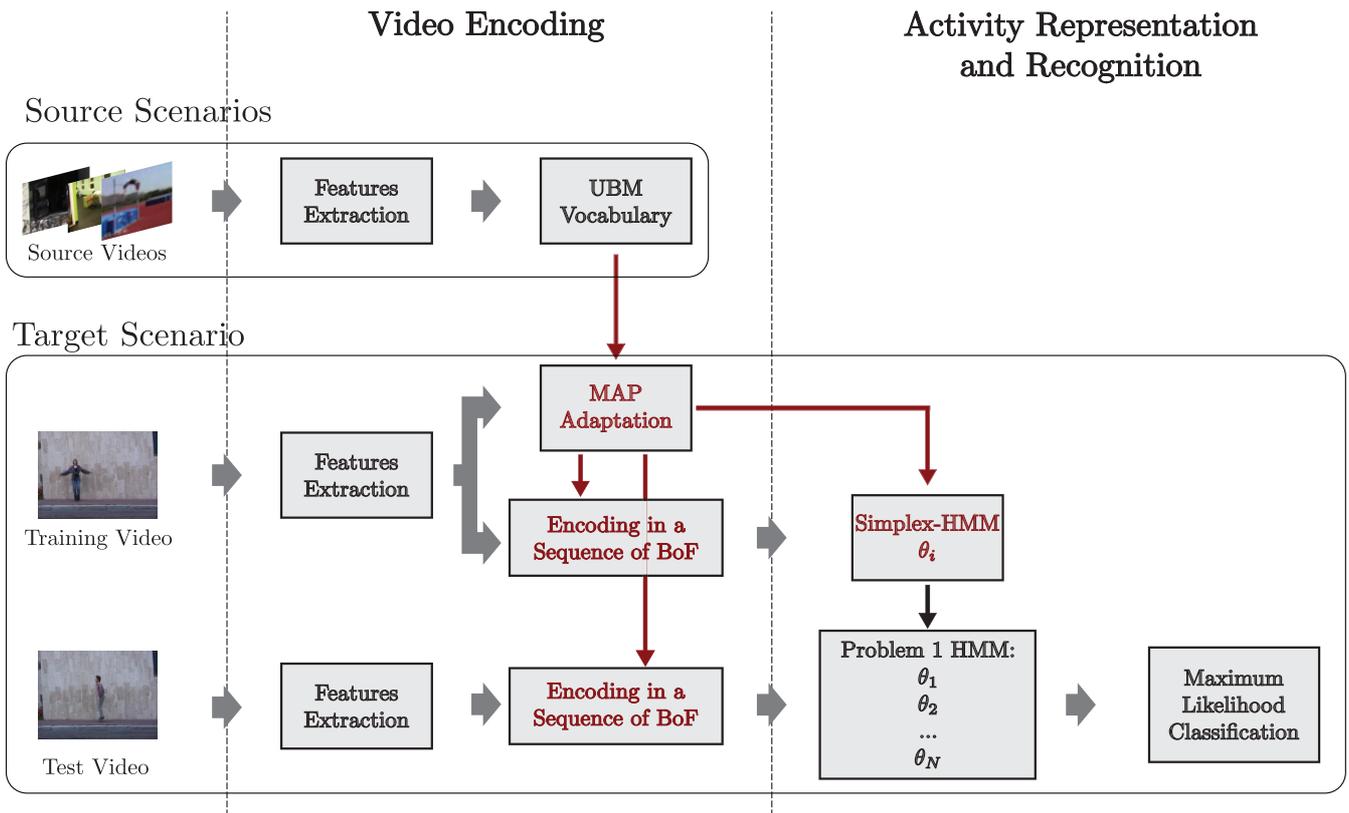


Fig. 1. Flow diagram of the proposed approach, highlighting in red the stages where the main novelties are introduced. *Video Encoding* is explained in Section III and *Activity Representation and Recognition* are explained in Section IV.

et al. [9], although they initially create a vocabulary with sequences of five classes, implying a relaxed model, they use a Transfer Learning stage to allow the inclusion of new classes without a retraining of the system in a hybrid approach.

In the most restricted scenario just one labelled sequence is available and its activity class should be trained from one example. To carry out with this restriction we propose in this paper the transfer learning of a Universal Background Model (UBM) [10] trained with extensive datasets available on-line, representing the vocabulary of a BoF model, to the target scenario where only a few labelled sequences are available, and training a specially designed case of a Hidden Markov Model (HMM).

A flow diagram of the proposed approach is depicted in Figure 1. From the wide range of features extractors available in the literature, Improved Dense Trajectories (IDT) [11] have shown state-of-the-art performance in several challenging datasets and so we use them in the Features Extraction stage. The IDTs are represented by the trajectory of spatio-temporal interest points during a window of time, the Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) [12] and the Motion Boundary Histograms (MBH) [13] features in the surrounding spatio-temporal cuboid of the trajectory. The proposed method extracts the IDTs from videos in public datasets of human activities, considered the source domain, and creates a UBM vocabulary modelled with a Gaussian Mixture Model (GMM) as done in [10], representing

general, person and scenario independent features. Unlike [10], the trained UBM represents a universe of features, and not a universe of activities (or speakers in their case). Once selected the target scenario, an initial labelled training video is recorded. The corresponding IDTs are extracted from this video and used in a twofold task. First, with the unordered IDTs, the UBM vocabulary is transferred to the target scenario using a Maximum a Posteriori (MAP) Adaptation, and obtaining a sequence specific vocabulary. Second, the IDTs are grouped into temporal windows where they are soft-assigned to the adapted vocabulary, obtaining a Bag of Features (BoF) per window. The BoF histogram is normalized so that it sums one, equivalent to say it belongs to a unit simplex. This way, the video is encoded as a sequence of BoF, and the activity is then modelled with a HMM which is a well known generative approach applied on time series.

The expectation-maximization (EM) algorithm is used to solve the difficult task of estimating a HMM. A complication with the EM algorithm is that there is in general no guarantee of reaching a global optimum, and local optima can at times be problematic, particularly with small training sets. Moreover, small training sets produce overfitting as we increase the number of parameters to train and the usual continuous HMM approach that incorporates GMM distributions may lead to an unstable EM algorithm [14]. In order to obtain a reliable system, the proposed Simplex-HMM is numerically stable, even with a single training sequence. Besides, the soft

assignment that leads to BoF seems more suitable than a hard assignment for the case of scarce training data.

Testing follows a process flow similar to training. First the IDTs are extracted and a sequence of BoFs is obtained using a temporal sliding window. The encoded video is then evaluated, given the HMM, as described in [15]. Maximum Likelihood Classification is used to identify the model that fits better the observation and therefore assign an action label. Using the defined one-shot learning system we have obtained state-of-the-art results in the public datasets Weizmann, KTH and IXMAS.

The two main contributions of this paper are summarised below:

- 1) We propose a new approach for video encoding based on Transfer Learning where a UBM vocabulary is obtained training a GMM with features from every sequence in the source domain, and adapted to the feature space extracted from the target scenario by a MAP adaptation of the GMM conforming a target domain vocabulary.
- 2) We define a HMM constrained to a sequence of vectors in a Simplex (Simplex-HMM), avoiding the numerical problems produced in the HMM training with scarce data.

The rest of the paper is divided in the following sections. In Section II a brief review of related research is given. In Section III the video encoding is explained, using the UBM adapted to the target domain. Section IV describes the Simplex-HMM. The experimental results are shown in Section V. Finally, a discussion of the work is done in Section VI.

II. RELATED WORK

A large volume of human activity recognition research has been performed in the last decades, mainly attempting to address two important questions: feature extraction and activity classification [16] [17] [18].

Many methods have been proposed for feature extraction, ranging from global to local descriptors. Global descriptors provide an holistic representation, while local descriptors usually use some encoding to reach that holistic representation. A single global descriptor encodes the activity by detecting a region of interest (ROI) through a bounding box or a contour where the person performs the activity. Examples of these approaches use human silhouettes to create Motion History Images (MHI) and Motion Energy Images (MEI) [19] [20], track the body contours creating spatio-temporal shapes [21] or obtain spatio-temporal volumes spanned by silhouette images [3]. Although they encode powerfully the activity information, they rely on accurate localization, background subtraction or tracking, being more sensitive to viewpoint, illumination changes, noise and occlusions. On the other side of the spectrum, the local descriptors are computed in small spatio-temporal volumes around interest points. Some descriptors only encode the image appearance [22], but usually spatio-temporal descriptors show better performance [23] [24] [12] [25]. The most common encoding of the extracted descriptors in a video activity is BoF representation, although recent state-of-the-art works have improved the results using Fisher

Vectors (FV) [26]. Methods based on local descriptors are robust to occlusion and are less dependent on viewpoint and illumination changes. However, they are too local and the encodings overlook important spatial and long-term temporal information. Recent approaches have obtain state-of-the-art results by expanding the local features to hybrid models where the spatio-temporal interest points are tracked during some frames obtaining a trajectory around which the descriptor is computed [27] [11].

Once the feature vectors are extracted, the recognition process becomes a classification problem. Several methodologies have been employed being discriminative or generative, considering or overlooking the temporal domain. A direct classification is performed selecting the class of the Nearest Neighbour (NN), the class of the closest training sequence. In order to avoid noise the k -NN methodology selects the most common label of the k closest sequences [20] [28]. Some methods consider the temporal domain to compute the distance, as Dynamic Time Warping (DTW) do [29] [30]. A widely used discriminative model is the Support Vector Machine (SVM) that learns an hyperplane in feature space that discriminates between two classes [4] [11]. SVM can be combined with BoF or FV losing the long-term temporal information in the encoding. On the other hand, some discriminative methods take into account this information as Conditional Random Fields (CRF) [31] [32] and its evolution Hidden CRF (HCRF) [33] do. The use of discriminative methods is not suitable for a strict one-shot learning approach as the method should be able to perform the training from a single example, so a generative method is needed. HMM is a generative method widely used in classification on time series [15] [34] [35].

The usual EM training process of HMM fails when training with a short number of examples and different strategies can be adopted in order to minimize the problem. Previous work has shown how a HMM modification called Fuzzy Discrete HMM (FDHMM) exploits a soft-assignment in a discrete HMM [36], obtaining a stable training with scarce data. Its application in activity recognition improves the performance [8] in a relaxed one-shot learning scenario, and the method can benefit from a Transfer Learning process [9]. In the last few years, transfer learning applied in human activity recognition has attracted the interest of researchers, as reflected in recent surveys [37] [38]. It is based in the use of external information from a source domain that complements the limited data available in a target domain where the recognition task is implemented. In [39] the authors use an inter-lingua in order to merge data from source domains and target domain, considering labelled data available only in the source domain, which classifies their method as uninformed supervised (US). In [40] a similar approach is implemented by creating a cross-domain codebook where labelled actions from both domains are modelled with BoF, being an informed supervised (IS) transfer learning method. Authors in [41] create a codebook with unlabelled data from the source domain and train the recognition with labelled data from the target domain, being an informed unsupervised (IU) transfer learning method. The literature is really extensive, but most of the methods use

similar approaches [42] [43]. The UBM vocabulary adaptation proposed in this paper lays in the US transfer learning as the source domain descriptors are labelled in a GMM training process and this GMM is later adapted to unlabelled target domain descriptors. A novelty introduced by this paper is the Transfer Learning per sequence, and specifically the adaptation of the UBM to as many vocabularies as labelled sequences in the target domain.

III. VIDEO ENCODING

Following the diagram depicted in the *Video Encoding* section of Figure 1, video encoding comprises two different tasks. Firstly, a UBM is modelled by a GMM using source videos, widely available, and afterwards, the adaptation of this UBM-GMM takes place on the target scenario. Secondly, as the video is encoded in a BoF taken into account a codebook, where the temporal information of the activity is lost, a temporal sliding window is used to recover this kind of information.

Next, we describe the temporal activity model by a GMM following a soft-assignment to a BoF approach. Later, we introduce the modification of the UBM vocabulary based in US transfer learning. Finally, we explain the motivation for using a sliding window to code temporal information given as a result a sequence of observations that will be used for training a HMM.

A. Soft-assignment-BoF

Over the past several years, many methods have modelled activities by encoding the extracted features in a single BoF, obtaining the codebook bins through a clustering algorithm of the training samples. Two of the most common clustering algorithms are k-means, which creates a Voronoi Tessellation, and GMM, optimized with an EM process. The former is defined only by the mean while the latter encodes second order information as includes both, the mean and the covariance and even a weight of the cluster. The proposed encoding uses the IDT features extracted from the activity videos and models the features space through a GMM. The number of clusters K can vary a lot in different approaches and empirically has been proven that a large number, in the order of thousands, is appropriate for BoF encoding [27], while FV allows a smaller number of Gaussians, in the order of hundreds [11].

From each video activity, a set of IDT feature vectors $\mathbf{Q} = \{\mathbf{q}_j\}$, $\mathbf{q}_j \in \mathbb{R}^D$ is extracted. Using the feature vectors of the training examples a GMM, $\lambda = \{\omega_i, \mu_i, \Sigma_i\}$, is calculated through an EM process. The general model of GMM supports full covariance matrices, but diagonal covariance matrices can satisfactorily approximate the original density modelling with a higher order GMM and they are computationally more efficient. Therefore, the framework uses diagonal covariance matrices and in addition it disregards GMM weights obtaining a simplified model $\lambda = \{\mu_i, \Sigma_i\}$.

Figure 2 shows a sample evaluated in every Gaussian of the GMM used to encode the data in a Soft-assignment-BoF. Using the simplified GMM, $\lambda = \{\mu_i, \Sigma_i\}$, the activity in a video is encoded with a BoF where each bin value v_{λ_i} is

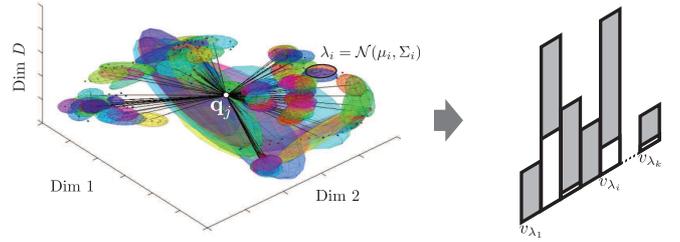


Fig. 2. GMM representation and Soft-assignment-BoF. Grey bars represents the BoF while white highlighted bars represents the \mathbf{q}_j sample contribution.

calculated by proportionally adding the contributions of every extracted feature \mathbf{q}_j to the specific Gaussian λ_i , as expressed in Equation 1, where M is the number of features extracted, and K is the number of Gaussians in the GMM.

$$v_{\lambda_i} = \frac{1}{M} \sum_{j=1}^M p(\lambda_i | \mathbf{q}_j) = \frac{1}{M} \sum_{j=1}^M \frac{\mathcal{N}(\mathbf{q}_j; \mu_i, \Sigma_i)}{\sum_{k=1}^K \mathcal{N}(\mathbf{q}_j; \mu_k, \Sigma_k)} \quad (1)$$

Thanks to the applied normalizations $\sum_{k=1}^K p(\lambda_k | \mathbf{q}_j) = 1$ every BoF vector belongs to the unit simplex $\Delta = \{\mathbf{v}_\lambda \in \mathbb{R}^K : v_{\lambda_i} \geq 0 : \sum_{k=1}^K v_{\lambda_k} = 1\}$.

With many feature samples, the proposed soft-assignment is unnecessary and the winner-takes-all rule usually applied in BoF approaches is sufficient. However, the one-shot learning objective is to obtain a representative model with only one activity example which contains few feature samples, therefore keeping as much information as possible, as a proper soft-assignment does, is essential.

B. Transfer Learning with MAP adaptation

In addition to the soft-assignment-BoF designed to deal with scarce data, another problem arises from the need of a codebook trained with the same scarce data, clearly insufficient using only one video activity example.

As mentioned before, although few samples are available in the target domain, plenty of videos can be obtained on-line as source domain from where the learning is transferred to the target domain. Several machine learning methods have used the approach of transfer learning from a source domain to a target domain. Especially successful have been the speaker verification systems based on UBM-GMM [10], although they transferred the class model and not a vocabulary. As previous research has proven, target domain information improves recognition system performance [44] [9].

Figure 3 represents the proposed Transfer Learning process showing a simplified 2D GMM trained in the source domain and adapted to the target domain. From the source domain a large number of IDTs is randomly extracted $\mathbf{S} = \{\mathbf{s}_j\}$, $\mathbf{s}_j \in \mathbb{R}^D$, being unlabelled data, and then used in a EM process to train the GMM which represents the UBM, $\lambda = \{\mu_i, \Sigma_i\}$. This UBM is later MAP adapted [10] to the target domain using only the available samples in this domain, they can be as few as the extracted from a single sequence, $\mathbf{Q} = \{\mathbf{q}_j\}$, $\mathbf{q}_j \in \mathbb{R}^D$. For Gaussian λ_i in the UBM, the probabilistic alignment of the feature vectors is computed with Equation 2.

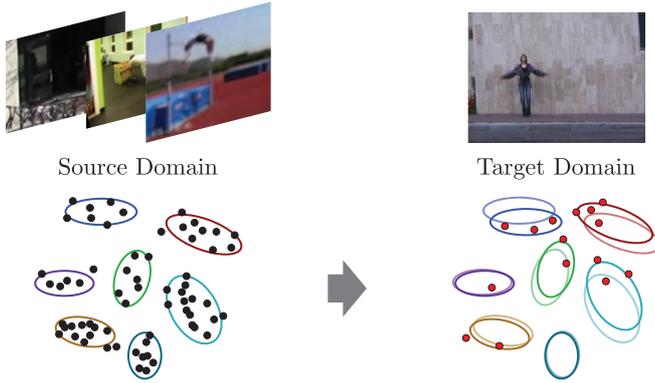


Fig. 3. Source GMM and MAP adapted GMM

$$p(\lambda_i | \mathbf{q}_j) = \frac{\mathcal{N}(\mathbf{q}_j; \mu_i, \Sigma_i)}{\sum_{k=1}^K \mathcal{N}(\mathbf{q}_j; \mu_k, \Sigma_k)} \quad (2)$$

These probabilistic alignments and the features vectors are used to compute the sufficient statistics of the mean with Equations 3 and 4. Weights have been disregarded so every Gaussian has the same weight, and covariance adaptation has been proven to be dispensable in most of the systems, so the system keeps the original covariance matrices in every new adapted GMM.

$$n_i = \sum_{j=1}^M p(\lambda_i | \mathbf{q}_j) \quad (3)$$

$$E_i(\mathbf{Q}) = \frac{1}{n_i} \sum_{j=1}^M p(\lambda_i | \mathbf{q}_j) \mathbf{q}_j \quad (4)$$

The sufficient statistics, computed with the target domain training data, are used to update the UBM, estimating the new means with Equation 5.

$$\hat{\mu}_i = \alpha_i E_i(\mathbf{Q}) + (1 - \alpha_i) \mu_i \quad (5)$$

The parameter α_i ($0 \leq \alpha_i \leq 1$) is an adaptation coefficient controlling the balance between old and new estimates that can also be estimated through Equation 6

$$\alpha_i = \frac{n_i}{n_i + rM} \quad (6)$$

where r is the controlling variable for adaptation and the rM term assures an equal adaptation independent on the number of IDT samples per example. After the MAP adaptation, a new GMM is obtained per video activity, $\hat{\lambda} = \{\hat{\mu}_i, \Sigma_i\}$, representing the new codebook used in the encoding.

C. Temporal Sliding Window

The former BoF video encoding loses the long-term temporal information of the activity, disregarding representative data. Therefore, a temporally windowed Soft-assignment-BoF encoding is proposed, shown in Figure 4. The MAP adaptation of the UBM vocabulary to the target scenario is performed with the features extracted from the whole video, as done

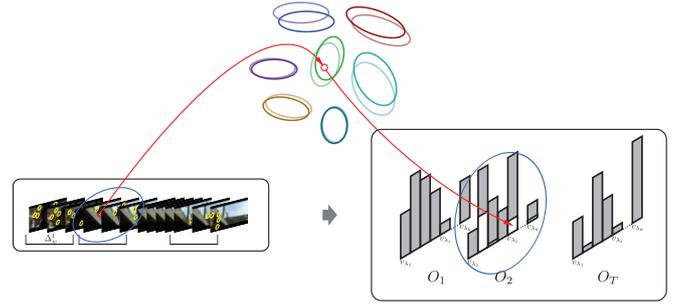


Fig. 4. BoF sequence of Windowed video

before. On the other hand, every N_w frames a new BoF is obtained, keeping the long-term temporal information in a sequence of BoF, $\mathcal{O} = \{O_1, \dots, O_T\}$. The encoding uses IDT as descriptors which are computed through a temporal window of length N_l , generally different to N_w . Each IDT, \mathbf{q}_j , has associated a temporal window Δ_l^j and influences proportionally to each Δ_w^t window given the Equation 7

$$\rho_{jt} = \frac{|\Delta_w^t \cap \Delta_l^j|}{N_w} \quad (7)$$

Each bin value, $v_{\lambda_i}^t$, associates to a specific BoF, O_t , is then calculated using Equation 8

$$v_{\lambda_i}^t = \frac{1}{\sum_{j=1}^M \rho_{jt}} \sum_{j=1}^M \frac{\rho_{jt} \mathcal{N}(\mathbf{q}_j; \hat{\mu}_i^t, \Sigma_i)}{\sum_{k=1}^K \mathcal{N}(\mathbf{q}_j; \hat{\mu}_k^t, \Sigma_k)} \quad (8)$$

Each of the Soft-assignment-BoF of the sequence belongs to the unit Simplex.

IV. ACTIVITY REPRESENTATION AND RECOGNITION USING SIMPLEX-HMM

Given an activity video, the proposed encoding represents the activity as a sequence of normalized BoF, $\mathcal{O} = \{O_1, \dots, O_T\}$, each one belonging to the simplex $\Delta = \{\mathbf{v}_\lambda \in \mathbb{R}^K : v_{\lambda_i} \geq 0 : \sum_{k=1}^K v_{\lambda_k} = 1\}$. These observations are \mathbb{R}^K vectors although the real dimensionality of the space is $(K - 1)$. In Figure 5a, a simplex of 3 dimensions is represented, and it can be observed that it is a triangle in a plane, so in reality it has only 2 dimensions. The sequence of normalized BoF can be used for training a classifier based in HMM, suitable for modelling time sequences. The *Activity Representation and Recognition* step shown in Figure 1 depicts the flow chart of a classifier based on a modified HMM, which is explained later in this section.

Formally, the parameters of the HMM are $\theta = \{N, A, B, \pi\}$. N is the number of states, i.e., $S = \{S_1, \dots, S_N\}$. Each observation, O_t is the emission produced by the hidden state z_t . The set of hidden states forms a sequence, $Z = \{z_1, \dots, z_T\}$ where $z_t \in S$. $A = \{a_{ij}\}$ is the state transition matrix where a_{ij} represents the transition probability from state i to state j , $a_{ij} = p(z_{t+1} = S_j | z_t = S_i)$. $\pi = \{\pi_i\}$ is the initial state probability distribution where $\pi_i = p(z_1 = S_i)$, $1 \leq i \leq N$ being S_i the state at the beginning of the time series. Finally, B represents

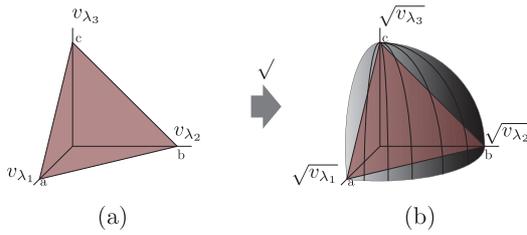


Fig. 5. (a) 3-dimensional simplex representation and (b) unit sphere portion encompassing the square root transformation of the simplex.

the observation probability distribution in every state where $b_j(O_t) = p(O_t | z_t = S_j)$. There are two main types of HMM based on the way the parameters in B are modelled: the discrete HMM for categorical observations, and the continuous HMM, where a Probability Density Function (PDF) is defined per state traditionally using a GMM. As previously mentioned, the observation space is the simplex Δ , which is a continuous space. Therefore the model should be a continuous HMM where the observations are vectors of real elements whose emission probability is traditionally approximated with a GMM per state.

In the simplex space, the emission probability can be modelled for instance with a Dirichlet distribution. However, the high dimensionality of the space causes numerical problems [45]. To exemplify these problems, we consider the simple case of a uniform distribution. In Equation 9 we observe how a uniform distribution would require the PDF $f(\mathbf{x}) = (K-1)!$, $\mathbf{x} \in \Delta$, so that the integral in the simplex is 1. So, with a high K any PDF is numerically infeasible.

$$\int_0^1 \int_0^{1-x_1} \dots \int_0^{1-\sum_{i=1}^{K-2} x_i} (K-1)! dx_{K-1} \dots dx_2 dx_1 = 1 \quad (9)$$

The emission probability can be modelled in the \mathbb{R}^K space, where there is no variance in the perpendicular dimension to the plane, or in a \mathbb{R}^{K-1} space obtained for instance by performing the Aitchison's solution to the compositional data. However, both cases suffer for the same problem produced when samples have a high dimensionality and their number is limited for training. The lack of available data produces overfitting of the parameters, and the high dimensionality of the data intensifies the problem, so the training of any sort of parameter related to covariance is hopeless. We have corroborated this performance with some preliminary experiments with the continuous HMM and one Gaussian per state, obtaining numerical problems as the likelihood has always gone to $-\infty$.

In this paper we propose to find a stable solution in spite of the high dimensionality. We simplify the observation model by defining the Euclidean distance between a mean vector and the observations, reducing the training parameters of the B function to only the means. Thus, we apply the observation model defined in Equation 10.

$$b_j(O_t) = e^{-\varphi \sqrt{\sum_{k=1}^K (v_{\lambda_k}^t - m_{jk})^2}} \quad (10)$$

Equation 10 shows φ and m_{jk} as free parameters but, as there are few samples in training and it is important to reduce as much as possible the parameters to learn, we experimentally fix the value of φ .

Although the function $b_j(O_t)$ is not a PDF it prevents the numerical problems while preserving the HMM properties thanks to the normalizations performed during the EM algorithm, as shown in the Equation 13 (see below). Moreover, the choice of the Euclidean distance in the exponential over a Gaussian comes from the experimental observation that a function highly peaked around data gives better results.

Considering that each $v_{\lambda_k}^t$ is an element in a histogram and represents the frequency of a specific feature model, it is possible to appreciate a drawback in the direct use of the Euclidean distance. Considering two normalized histograms of the same dimensionality, $A = \{a_i\}$ and $B = \{b_i\}$, $1 \leq i \leq N$ and $\sum_{i=1}^N a_i = \sum_{i=1}^N b_i = 1$, if $a_j = 1$, so $a_i = 0, \forall i \neq j$, and $b_j = 0$, then the dissimilarity between A and B should be maximum independent on the values of $b_i, \forall i \neq j$, and then the distance from A to B should remain constant for all values of $b_i, \forall i \neq j$. In Figure 5a this distance would be represented by the distance from point a to any point in segment bc , which is not constant and the middle point of the segment is closer to a than the edges, being more significant with high dimensionality. To tackle this drawback, we propose to replace the Euclidean distance with the Hellinger distance, which is equivalent to transforming the points in the simplex to a portion of a hypersphere of unit radius by applying the square root of the vector element ($v_{\lambda_k}^t \rightarrow \sqrt{v_{\lambda_k}^t}$), as shown in Figure 5b. However, for the sake of simplicity we do not impose the condition that $\sum_{k=1}^K m_{jk} = 1$ in the optimization process. For the sake of clarity the transformed $\sqrt{v_{\lambda_k}^t}$ is not mentioned in the following equations because the formulation is equivalent to using v_{λ_k} , just changing one for the other and then the formulation is valid in both cases.

Given on, or several training observations, the HMM parameters can be estimated using the Maximum Likelihood through a Baum-Welch algorithm. This iterative estimation is obtained by maximizing the Baum's auxiliary function $Q(\hat{\theta}, \theta)$ [15] [46].

$$Q(\hat{\theta}, \theta) = \sum_Z p(Z|\mathcal{O}, \theta) \ln p(\mathcal{O}, Z|\hat{\theta}) \quad (11)$$

Defining $\gamma_t(i) = p(z_t = S_i|\mathcal{O}, \theta)$ and $\xi_t(i, j) = p(z_t = S_i, z_{t+1} = S_j|\mathcal{O}, \theta)$, the function Q can be expressed as:

$$Q(\hat{\theta}, \theta) = \sum_{j=1}^N \gamma_1(j) \ln \pi_j + \sum_{t=1}^{T-1} \sum_{i=1}^N \sum_{j=1}^N \xi_t(i, j) \ln a_{ij} + \sum_{t=1}^T \sum_{j=1}^N \gamma_t(j) \ln(b_j(O_t)) \quad (12)$$

The EM algorithm followed requires a modification in the M step to optimize the function with respect to the proposed $b_j(O_t)$, which prevents from the problems arisen with scarce data. This special case is presented below.

E-step: This step implies the calculation of functions $\xi_t(i, j)$ and $\gamma_t(i)$. They are calculated with the standard equations for a general observation model $b_j(O_t)$:

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad (13)$$

where $\alpha_t(i) = p(O, z_t = S_i | \theta)$ and $\beta_t(j) = p(O | \theta, z_t = S_j)$ are auxiliary probabilities calculated by the forward and backward algorithms [15]. Equation 13 is computed using the proposed $b_j(O_{t+1})$.

Finally, $\gamma_t(i)$ is obtained with Equation 14.

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j) \quad (14)$$

M-step: This process calculates the $\hat{\pi}_i$, \hat{a}_{ij} and \hat{m}_{jk} that maximize $Q(\hat{\theta}, \theta)$. The optimizations of $\hat{\pi}_i$, \hat{a}_{ij} and \hat{m}_{jk} are obtained by maximizing Equation 12. In our method the optimizations of $\hat{\pi}_i$, \hat{a}_{ij} are the same as in the traditional method but \hat{m}_{jk} should be computed differently as its term changes in the mentioned equation. Therefore, Equation 15 has to be maximized with respect to m_{jk} :

$$\sum_t \sum_j \gamma_t(j) \ln(b_j(O_t)) = \sum_t \sum_j \gamma_t(j) \left(-\varphi \sqrt{\sum_{k=1}^K (v_{\lambda_k}^t - m_{jk})^2} \right) \quad (15)$$

By setting $\frac{\partial}{\partial m_{jk}} = 0$, the following equation is obtained:

$$\varphi \sum_{t=1}^T \gamma_t(j) \frac{(v_{\lambda_k}^t - m_{jk})}{\sqrt{\sum_{k'=1}^K (v_{\lambda_{k'}}^t - m_{jk'})^2}} = 0 \quad (16)$$

Since m_{jk} does not depend on t and $\gamma_t(j)$ are treated as constants in the M-step once computed in the E-step, equation 17 can be easily derived:

$$m_{jk} = \frac{\sum_{t=1}^T \gamma_t(j) \frac{v_{\lambda_k}^t}{\sqrt{\sum_{k'=1}^K (v_{\lambda_{k'}}^t - m_{jk'})^2}}}{\sum_{t=1}^T \gamma_t(j) \frac{1}{\sqrt{\sum_{k'=1}^K (v_{\lambda_{k'}}^t - m_{jk'})^2}}} \quad (17)$$

Then, since m_{jk} is on the left and on the right side, the equation is solved by a fixed point iteration, obtaining \hat{m}_{jk} when convergence is achieved.

Each Simplex-HMM trained with a single video has a specific GMM computed with the MAP adaptation, so the Simplex-HMM model is defined by $\Gamma(A, B, \pi, \hat{\mu}_i, \Sigma_i)$. As each training sequence is modelled with a Simplex-HMM the inclusion of new training sequences implies a linear increment on the storage space and the required computational power for testing.

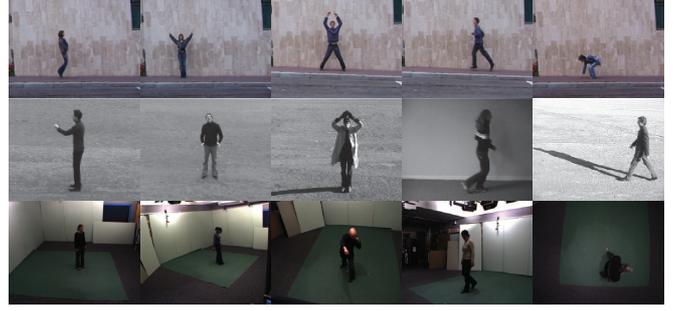


Fig. 6. Target Domain Datasets: Weizmann (1st row), KTH (2nd row) and IXMAS (3rd row)

V. EXPERIMENTS AND RESULTS

A. Datasets

This paper focuses on human activity recognition applied on constrained scenarios, where videos are obtained by fixed viewpoint cameras. The proposed algorithm is trained using human motion information from external video sources using MAP adaptation, as described in Section III-B. Our method is evaluated using several datasets that accomplish the source and target domain constraints. We have selected three source domain datasets that include a high variability in unconstrained video clips that simulate the easily obtainable ones from the Internet. On the other hand, we have selected three popular datasets in the human activity recognition field as target domain where the videos are recorded from fixed cameras. Additionally, a dataset with unconstrained video recording has been selected as target dataset in order to evaluate the performance of the algorithm in general purpose applications.

Source Domain Datasets Three public and extensive datasets, HMDB51 [1], OlympicSprots [2] and Virat Release 2.0 [47], are used as source domain. They include a high variability of movements in several locations. The three datasets combined have 79 different activity classes extracted from Youtube, movies or surveillance cameras in 7878 video clips.

Target Domain Datasets The Weizmann dataset [3] is composed by 93 low-resolution (180 x 144, 50 fps) video sequences showing nine different people, each performing 10 natural activities: *bend*, *jumping-jack*, *jump-forward-on-two-legs*, *jump-in-place-on-two-legs*, *run*, *gallop-side-ways*, *skip*, *walk*, *wave-one-hand* and *wave-two-hands*. The IXMAS dataset [5] is composed by 5 camera viewpoints (390 x 291, 23 fps) of 11 actors performing 3 times each of the 13 activities included: *check-watch*, *cross-arms*, *scratch-head*, *sit-down*, *get-up*, *turn-around*, *walk*, *wave*, *punch*, *kick*, *point*, *pick-up* and *throw*. The KTH dataset [4] has been captured in 4 different scenarios where static cameras have recorded, at low-resolution (160 x 120, 25 fps), 25 subjects performing several times six types of activities: *walking*, *jogging*, *running*, *boxing*, *hand-waving* and *hand-clapping*. Frame examples of these datasets are shown in Figure 6. Finally, we have selected the UCF11 dataset [48] composed by unconstrained video clips of 11 categories obtained from YouTube.

All videos are processed by means of the state-of-the-art

IDT¹ extractor. From the IDTs extracted from the Source Domain datasets, 100000 are randomly selected and used for the GMM training, obtaining 5000 Gaussians, which represent the UBM vocabulary. Below we detail the performed experiments using two approaches: Strict One-shot learning and Relaxed One-shot learning.

B. Strict One-shot learning

Considering the proposed activity representation, there are two possible ways of modelling an activity. First, the simplest and fastest method uses the original UBM trained with the source datasets, which is represented by a GMM $\lambda = \{\mu_i, \Sigma_i\}$. A Simplex-HMM (SHMM) is then trained per example obtaining the model $\Gamma(A, B, \pi, \mu_i, \Sigma_i)$ where μ_i and Σ_i is shared in all models. We define the Simplex-HMM as an ergodic two-state model and set $\varphi = 1.5$, as experiments shown that the performance is rather insensitive when φ is in the range (1, 2). The second approach, on the other hand, adapts the UBM vocabulary to an improved GMM per example using the explained MAP adaptation. In [10] they suggest the insensitivity of the method to the parameter r that we experimentally corroborate, selecting finally $r = 0.014$, where only the mean is modified. Again, a Simplex-HMM is trained per activity example, but each activity model is represented in the adapted UBM vocabulary, which implies different GMM means per example as specific information and only Σ_i is shared among models, $\hat{\Gamma}(A, B, \pi, \hat{\mu}_i, \Sigma_i)$. This second approach is called MAP+SHMM in the experiments.

In addition to the previous representations, a special instance of the SHMM approach is performed in Weizmann dataset in order to validate the proposed algorithms. If the source dataset used to train the UBM is the whole target domain, including both, train and test examples, then the obtained UBM is the optimal that can be reached with the method configuration, so this special case is considered the UBM ground truth and is labelled with the name Opt+SHMM in the experiments. It is worth noting that in a real world application this is infeasible and the Opt+SHMM configuration is only used to represent the experimental ceiling of our methodology.

We perform the following initialization for the EM algorithm used in the Simplex-HMM with 2 states: $\pi_1 = 1$ and $\pi_2 = 0$, the transition matrix is randomly initialized, and finally the initial mean vectors, m_1 and m_2 , are the observations closest to times $\frac{T-1}{4} + 1$ and $\frac{3(T-1)}{4} + 1$ of the training sequence.

The experiments in this section are conducted as follows. Using one-subject-out model, num training sequences per class are randomly selected from the remaining subjects. The value of num goes from 1 to the maximum available sequences. The result per subject are the average of 100 runs, and the final result is the average of all subjects.

Figure 7 shows on the left a graph with the performance of the method proposed in this paper using the three activity representations described previously. It is worth noting the Opt+SHMM results because they justify the suitability of the

selected activity representation, i.e. the IDT features encoded in a sequence of BoFs which trains a Simplex-HMM. The results are impressive with only one sequence of each class, as it reaches a 91.8%, and when using the 8 sequences available it almost reaches the 100%, which is comparable to the state-of-the-art results. However, as explained before, this is only possible if the feature space is properly modelled with a GMM, and in this case we have used information from the whole dataset. We can conclude from this result that a proper feature space representation becomes an important objective to be achieved. Thus, we have trained a UBM vocabulary, represented with a GMM of IDTs, obtained from a set of videos as diverse as possible (using the three source domain datasets). Applying the SHMM configuration with the GMM trained with three source datasets, the method performance falls significantly against the Opt+SHMM configuration, although it still obtains a satisfactory 80.11% using only one sequence per class. Finally, the results are improved by adapting the GMM to the scenario, but as each model uses a single example available, the MAP adaptation has to be performed to this limited available data using the MAP+SHMM configuration. Figure 7 demonstrates how this adaptation improves the results during all the series, which implies that the adaptation to the target scenario improves the features representation. On the right of Figure 7, the Confusion Matrix of the MAP+SHMM method using only one sequence for training is depicted. From all the classes the greatest confusion is produced among activities that involve subject displacement (*jump*, *run*, *side*, *skip*), caused by the model giving more importance to the displacement information than the limbs movements. Some works as the one in [3] incorporates a preprocessing that compensates the displacement, but because of the use of background subtraction and its complications in some scenarios we have opted to avoid it.

Figure 8 shows the performance of our proposed method on KTH dataset, only including results for SHMM and MAP+SHMM representations. In this case there are up to 70 examples for training, and as the graph shows there is a constant gap between SHMM and MAP+SHMM almost independent on the training examples, which clearly demonstrates the improvement obtained using the proposed adaptation to the scenario. Attending to the Confusion Matrix we can observe the same phenomenon produced in Weizmann dataset, the classes with displacement (*walk*, *jog*, *run*) are mainly confused among them. Additionally, we observe in this dataset how the “static” classes (*boxing*, *wave*, *clap*) are confused as well among them. Unlike Weizmann, the “static” activities in KTH are all described by the movement of the subject arms, which results to the difficulty to distinguish them.

Finally, we repeat the experiments using the IXMAS dataset but avoiding the *point* and *throw* activities as suggested in [5] and Figure 9 shows these results. It is worth noting that IXMAS dataset is recorded simultaneously with 5 fixed cameras each one with different viewpoint. We conduct the experiment separately per camera and the shown results is the average of these experiments. Due to the free subject position in the scenario some activities, like *check-watch*, *cross-arms* but also others, are occluded by the subject body in some

¹IDT descriptor code can be downloaded in http://lear.inrialpes.fr/people/wang/download/improved_trajectory_release.tar.gz

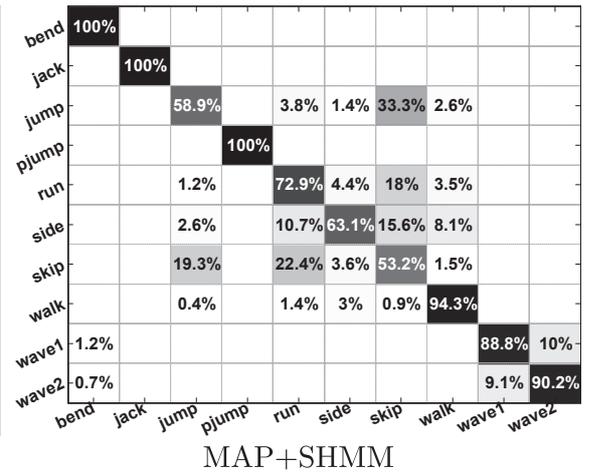
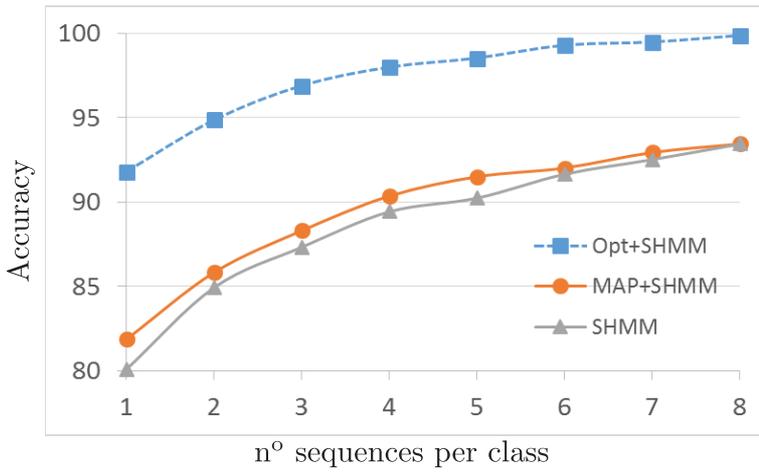


Fig. 7. Strict One-shot learning results in Weizmann dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+SHMM method using only one example per class in training

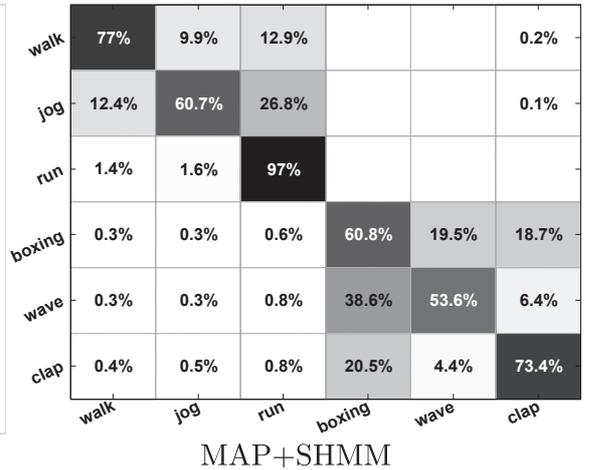
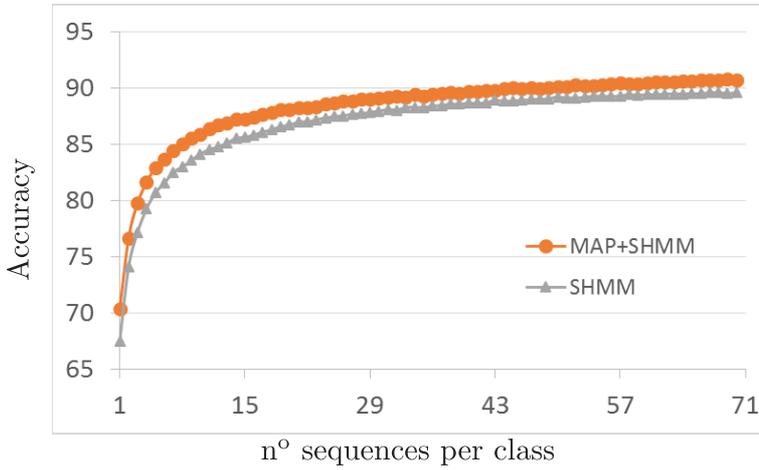


Fig. 8. Strict One-shot learning results in KTH dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+SHMM method using only one example per class in training

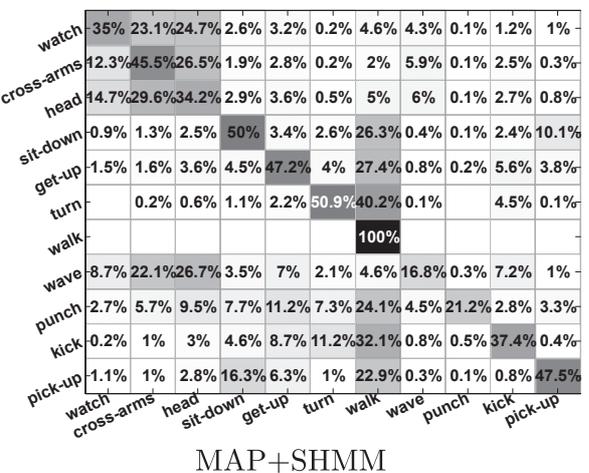
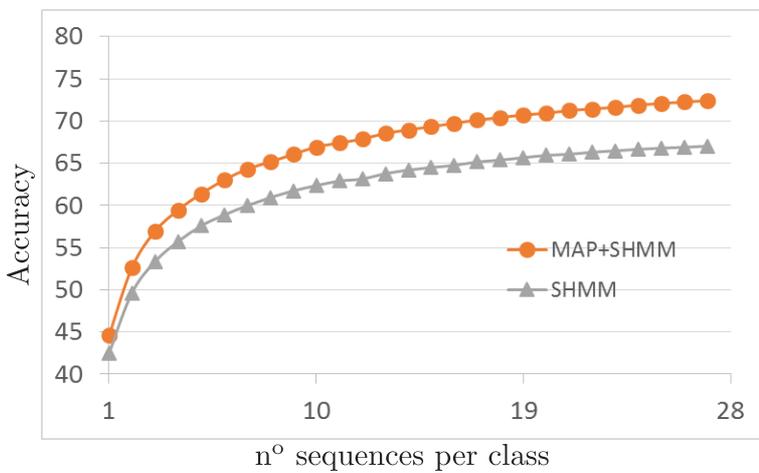


Fig. 9. Strict One-shot learning results in IXMAS dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+SHMM method using only one example per class in training

cameras generating a worse performance in comparison to the other datasets. However, we ratify the suitability of the MAP Adaptation in the graph. The Confusion Matrix highlights two phenomena: First, the *walk* activity is always right but several other activities are confused with it, which indicates a bias in that model that could be compensated but potential solution is beyond the scope of this paper. Second, there are four activities (*check-watch*, *cross-arms*, *scratch-head* and *wave*) that are mainly confused among them. The lower amount of movement in these activities contributes to a higher dependence in the camera viewpoint and therefore their confusion.

As the experiments in IXMAS have been performed per camera we can evaluate them separately. The results of each camera accuracy is shown in Table I using only one example per training class or using the maximum of 27. In addition to the MAP+SHMM improvement, we can highlight that this improvement is more significant in camera 5, which is a zenith camera. The viewpoint of this camera is rarely used and then the movements are worse represented in the UBM. Thus, the MAP Adaptation produces a greater impact.

We have found only one paper covering the Strict One-shot learning paradigm [6], although indirectly. They propose a sequence representation based in their defined local space-time descriptors. Afterwards, computing a distance among sequences they select the class of the closest one, being a Strict One-shot learning as the representation do not need information of other sequences. The experimental results are presented by two graphs, one for Weizmann and other for KTH, that show the performance obtained using only one sequence per class in the training. In Table II we compare our results with this work, extracting their results approximately from the graphs, showing clearly how our approach improves the previous results significantly in the two comparable datasets. In IXMAS we compare with [9], where silhouettes were used as descriptors. In that work, only camera1 was used, starting with 5 sequences and not only one, although they studied the performance including new classes, which approximates the method to Strict One-shot learning. The proposed method clearly overcame their results using the five cameras (camera5 gives the worst results), and learning from the first sequence. Moreover, in Section II we have mentioned FDHMM [36] which is able to train a stable HMM from a single sequence of vectors in a unit simplex, as our proposal does. However, the histogram distribution is not considered and therefore it fails in the experiments as shown in Table II where we conduct the FDHMM experiments with and without the MAP adaptation.

After proving the suitability of the method for constrained scenarios, we evaluate the Strict One-shot learning in the unconstrained dataset UCF11. We randomly chose 1, 10, 20, 30, 40 and 50 video-clips of each class for training and we test the rest. The average of 50 runs is the result shown in Table III. This experiment is comparable to the results found in [7]. Initially, when choosing 1 sequence for training, we compare SHMM with MAP-SHMM, obtaining similar results (30.67% and 30.83% respectively). As sequences belong to unconstrained scenarios the UBM adaptation does not improve the accuracy and we discard it. On the other hand, we verify how our approach has a good performance compared with [7].

C. Relaxed One-shot learning

As the literature is scarce in One-shot learning methods for human activity recognition, we add a new experiment using the Relaxed One-shot learning methodology. In this experiment we select the video examples of one subject, and from them only one example per class. These examples are used for training one Simplex-HMM per class, and all the other subjects are used in testing. The Relaxed One-shot learning allows us to apply the MAP adaptation to the features extracted from all the training examples, which implies a better adaptation in comparison to the previous experiments. Moreover, the GMM is now shared among the Simplex-HMMs as in the SHMM method. However, this process has some constraints in comparison to the Strict methodology as a sequence per class is necessary from the beginning, which initially can be expensive to obtain, and implies a less flexible addition of new examples and classes.

Table IV shows the results in Relaxed One-shot learning using our method in comparison with some results found in the literature. In [7], spatio-temporal sub-actions based on optical flow are defined and modelled from all the sequences available for training. In [8], the descriptors are Motion History Images (MHI) [19] based on silhouettes computed in a fixed temporal window. MHIs from all the sequences in training are used to model the feature space. Both methods are outperformed, demonstrating the suitability of the proposed method. Again, it is shown the improvement achieved by the use of the MAP Adaptation to the scenario. The improvement of these results compared to the Strict methodology happens not only because of the adaptation, as the SHMM method does not benefit from it, but also because every Simplex-HMM uses the same actor, and therefore the difference is not on the actor features but only on the activity.

In Figures (7, 8 and 9) we have shown the evolution of the method performance while increasing the number of labelled examples. However, when many training sequences are available the MAP Adaptation to each sequence is not optimal as there is enough information for training a specific GMM. Therefore, using a leave-one-person-out cross validation methodology we have carried out a new experiment. Obtaining a GMM per person with the examples of the remaining subjects, making unnecessary any adaptation, we have trained a SHMM per example, keeping unchanged the rest of the system. We call this new experiment OneOut-SHMM. Our method achieves almost state-of-the-art results in Weizmann and KTH, being the gap of Weizmann caused by only one sequence misclassified. However, in IXMAS we obtain worse results, probably partially caused by a naive fusion of cameras, we have selected the class of the HMM producing the highest likelihood among all cameras. These results show how the method, although performing reasonably well with several training data, it is not the best suited for this task.

VI. CONCLUSION

The human activity representation proposed in this paper has been proved suitable in activity recognition in fixed-background constrained scenarios with very few available

TABLE I
ACCURACY USING STRICT ONE-SHOT LEARNING IN EACH OF THE IXMAS CAMERAS USING 1 AND 27 EXAMPLES FOR TRAINING.

	camera1		camera2		camera3		camera4		camera5	
	1	27	1	27	1	27	1	27	1	27
SHMM	45.7	73.6	44.5	71.5	43.1	69.7	48.5	69.7	30.6	50.6
MAP+SHMM	48.3	78.5	45.8	76.1	44.7	71.5	50	74.8	34	61.2

TABLE II
STRICT ONE-SHOT LEARNING WITH ONE EXAMPLE PER CLASS.

	Weizmann	KTH	IXMAS
FDHMM [36]	68.17%	67.16%	25.36%
MAP+FDHMM [36]	69.61%	67.6%	33.7%
SHMM	80.11%	67.53%	42.47%
MAP+SHMM	81.88%	70.39%	44.58%
<i>Seo and Milanfar</i> [6]	75%	65%	-
<i>Rodriguez et al.</i> [9]	-	-	35%*

*Not directly comparable as they use only one camera and 5 initial sequences, which is less restrictive.
(-) Lack of results in the referenced papers.

TABLE III
STRICT ONE-SHOT LEARNING FOR UCF11 DATASET. RESULTS FOR OUR APPROACH COMPARED WITH [7].

	1	10	20	30	40	50
SHMM	30.7%	59.2%	69%	75%	79.2%	82%
<i>Yang et al.</i> [7]	19.3%	31.3%	39.2%	46.3%	50%	51%

training examples. The introduced Simplex-HMM facilitates the modelling of an activity using limited amount of data, as few as one example per class, thanks to the reduction of parameters to train, exploiting the proposed Simplex constraints of the samples. In the Weizmann experiments we have seen how these representation obtains great results if the feature space is properly modelled with a GMM.

In order to obtain the best representation of the feature space

TABLE IV
RELAXED ONE-SHOT LEARNING WITH ONE EXAMPLE PER CLASS. RESULTS OF OUR APPROACH AND TWO STATE-OF-THE-ART METHODS.

	Weizmann	KTH	IXMAS
SHMM	84.18%	76.65%	52.84%
MAP+SHMM	87.12%	80.21%	56.43%
<i>Yang et al.</i> [7]	80%	-	-
<i>Orrite et al.</i> [8]	81.1%	-	-

(-) Lack of results in the referenced papers.

TABLE V
ACCURACY OF THE PROPOSED METHOD WITH SEVERAL TRAINING EXAMPLES, IN COMPARISON WITH STATE-OF-THE-ART RESULTS.

	Weizmann	KTH	IXMAS
OneOut+SHMM	98.9%	94.2%	82.5%
<i>Liu et al.</i> [49]	100%	94.8%	95.5%

using a GMM, but taking into account the limited data available in the target scenario, we propose an adaptation of a UBM trained in source datasets. We have proved how an adaptation of the UBM to the target domain information modelled only with one example in the target scenario improves the results obtained using directly the original UBM. The results obtained in the Transfer Learning stage demonstrate the value of the proposed method adopted, but also that there is still scope of improvement in future work.

It is worth noting that the proposed algorithm assumes a limited number of available labelled sequences from the target dataset. However, the more labelled sequences, the more storage space and computational power is required for inferring sequences. Therefore, in cases where more data is available, other existing algorithms may be more practical and effective as shown in the experiments.

Also, it is worth noting the shortage of relevant work in the literature existing in One-shot learning for human activities. Especially in the Strict One-shot learning paradigm very little work has been done, and our point of view is that it is the most appropriate in many real life applications, thanks to its flexibility including new examples and classes. Although the human activity recognition community is tending to focus in large unconstrained datasets, more research in this field can accelerate the installation of recognition systems in new scenarios.

REFERENCES

- [1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [2] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 392–405.
- [3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.
- [4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *International Conference on Pattern Recognition*, 2004. [Online]. Available: <http://www.nada.kth.se/cvap/actions/>
- [5] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006. [Online]. Available: <http://4drepository.inrialpes.fr/public/viewgroup/6>
- [6] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 867–882, 2011. [Online]. Available: <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.156>
- [7] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, Jul. 2013.
- [8] C. Orrite, M. Rodriguez, and M. Montañes, "One-sequence learning of human actions." in *Human Behavior Understanding*, A. Salah and B. Lepri, Eds., vol. 7065. Springer Berlin / Heidelberg, 2011, pp. 40–51.

- [9] M. Rodriguez, C. Medrano, E. Herrero, and C. Orrite, "Transfer learning of human poses for action recognition," in *Human Behavior Understanding*, 2013.
- [10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
- [11] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
- [12] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2008.
- [13] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision (ECCV)*, 2006. [Online]. Available: <http://lear.inrialpes.fr/pubs/2006/DTS06>
- [14] T. Shinozaki and M. Ostendorf, "Cross-validation and aggregated {EM} training for robust parameter estimation," *Computer Speech & Language*, vol. 22, no. 2, pp. 185 – 195, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0885230807000472>
- [15] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, feb 1989.
- [16] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, Nov 2008.
- [17] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, Jun. 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.imavis.2009.11.014>
- [18] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, Feb. 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.cviu.2010.10.002>
- [19] A. Bobick and J. Davis, "Real-time recognition of activity using temporal templates," in *Applications of Computer Vision, 1996. WACV '96., Proceedings 3rd IEEE Workshop on*, Dec 1996, pp. 39–42.
- [20] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001. [Online]. Available: <http://dx.doi.org/10.1109/34.910878>
- [21] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2005, pp. 984–989.
- [22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: <http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>
- [23] I. Laptev, "On space-time interest points," *International Journal of Computer Vision (IJCV)*, vol. 64, pp. 107–123, 2005.
- [24] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 14th International Conference on Computer Communications and Networks*, ser. ICCCN '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 65–72. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1259587.1259830>
- [25] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision (ECCV)*, 2012.
- [26] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1817–1824.
- [27] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision (IJCV)*, vol. 103, pp. 60–79, 2013.
- [28] D. Batra, T. Chen, and R. Sukthankar, "Space-time shapelets for action recognition," in *Motion and Video Computing, 2008. WMVC 2008. IEEE Workshop on*, Jan 2008, pp. 1–6.
- [29] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1896–1909, Dec 2005.
- [30] B. Yao and S.-C. Zhu, "Learning deformable action templates from cluttered videos," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 1507–1514.
- [31] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 210–220, Nov. 2006. [Online]. Available: <http://dx.doi.org/10.1016/j.cviu.2006.07.014>
- [32] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: <http://dl.acm.org/citation.cfm?id=645530.655813>
- [33] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.
- [34] X. Feng and P. Perona, "Human action recognition by sequence of wavelet codewords," in *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, 2002, pp. 717–721.
- [35] W.-L. Lu and J. Little, "Simultaneous tracking and action recognition using the pca-hog descriptor," in *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, June 2006, pp. 6–6.
- [36] H. Uğuz, A. Öztürk, R. Saraçoğlu, and A. Arslan, "A biomedical system based on fuzzy discrete hidden markov model for the diagnosis of the brain diseases," *Expert Systems with Applications*, vol. 35, no. 3, pp. 1104–1114, Oct. 2008. [Online]. Available: <http://dx.doi.org/10.1016/j.eswa.2007.08.006>
- [37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, October 2010.
- [38] D. Cook, K. Feuz, and N. Krishnan, "Transfer learning for activity recognition: a survey," *Knowledge and Information Systems*, pp. 1–20, 2013. [Online]. Available: <http://dx.doi.org/10.1007/s10115-013-0665-3>
- [39] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3209–3216. [Online]. Available: <http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#LiuSKS11>
- [40] W. Bian, D. Tao, and Y. Rui, "Cross-domain human action recognition," *IEEE Transactions on Systems, Man, and Cybernetics. B Cybernetics*, vol. 42, no. 2, pp. 298–307, 2012.
- [41] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Asian Conference on Computer Vision (ACCV)*, 2011.
- [42] D. H. Hu, V. W. Zheng, and Q. Yang, "Cross-domain activity recognition via transfer learning," *Pervasive and Mobile Computing*, vol. 7, pp. 344–358, June 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.pmcj.2010.11.005>
- [43] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 594–611, 2006.
- [44] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, 2010, pp. 1998–2005. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2010.5539875>
- [45] T. P. Minka, "Estimating a Dirichlet distribution," Tech. Rep., 2009. [Online]. Available: <http://research.microsoft.com/~{minka>
- [46] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [47] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2011, pp. 3153–3160.
- [48] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009. [Online]. Available: http://server.cs.ucf.edu/~vision/projects/liujg/realistic_action_recognition.html
- [49] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1860–1870, Dec 2013.