Human Action Recognition via Multi-view Learning

Tianzhu Zhang\textsuperscript{1,2}, Si Liu\textsuperscript{1,2}, Changsheng Xu\textsuperscript{1,2}, Hanqing Lu\textsuperscript{1,2}
\textsuperscript{1} National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China
\textsuperscript{2} China-Singapore Institute of Digital Media, Singapore 119613, Singapore
\{tzzhang, sliu, csxu, luhq\}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we propose a novel approach to automatically learn a compact and yet discriminative representation for human action recognition. Considering the static visual information and motion information, each frame is represented in two feature subsets (views) and Gaussian Mixture Model (GMM) is adopted to model the distributions of those features. In order to complement the strengths of the different features (views), a Co-EM based multi-view learning framework is introduced to estimate the parameters of GMM instead of conventional single view based EM. Then Gaussian components are considered as video words to describe videos with different time resolutions. Compared with the traditional method to recognize action, there are several advantages with the proposed method using Co-EM strategy. First, complex actions are efficiently modeled by GMM, and the number of its component is automatically determined with the Minimum Description Length (MDL). Second, because the imperfectness of single view can be compensated by the other view in the Co-EM, the resulting bag of video words are superior to that formed by any single view. To the best of our knowledge, we are the first to try the Co-EM based multi-view learning method for action recognition and obtain significantly better results. We extensively verify our proposed approach on two publicly available challenging datasets: the KTH dataset and Weizmann dataset. The experimental results show the validity of our proposed method.

Keywords
Co-EM, Action Recognition, GMM, Bag of Words, Multi-view Learning.

1. INTRODUCTION

With the explosive growth of multimedia content on broadcast and Internet, it is urgently required to make the unstructured multimedia data accessible and searchable with great ease and flexibility. Action recognition is particularly crucial to understanding video semantic concepts for video summarization, indexing and retrieval purposes. There are two key issues in the video based action recognition in practice. One is the feature extraction and fusion. There are various features derived from significantly different modalities, such as static visual cues, e.g. shape and appearance, as well as dynamic cues, e.g. spatial-temporal trajectories and motion field. Such diversity raises the question of the relative importance of these sources and also to what degree they compensate for each other. Because these different kinds of features are heterogeneous, it is difficult to mine their effectiveness just with simple fusion. The other is how to model different actions and measure their similarities for recognition. As we know, different actions have different time resolutions, and may have some similar elements. In addition, though they are the same actions, they may contain some different elements. Therefore, it is difficult to recognize different actions.

Static visual information gives very strong cues on activities. That is, in many cases, humans are able to recognize many actions from a single image (see for instance Figure 1(a)). A set of key poses can represent an action, and the sets of key poses show differences from action to action, though with possible partial overlap among activities. Therefore, Daniel Weinland and Edmond Boyer [18] have successfully recognize different actions using key poses. But such methods ignore temporal information between poses. As an intuitive example, they cannot tell reverse pairs of actions such as sitting down and standing up. Since motion is an important source of information for classifying human actions. A popular approach adopted by vision researchers for action recognition is to utilize the motion of human actor, where the motion is quantified in terms of the optical flow computed from the sequence of images depicting the action [2]. Therefore, the low-level motion features, i.e. optical flow feature, can be used to describe the temporal dependencies. But different actions may contain similar optical flows, which make the optical flow feature less discriminative. To sum up, any single feature is not sufficient for action recognition. So we utilize both of the views to complement each other and enhance their discriminativeness.

However, how to fuse and associate these heterogeneous features

![Figure 1: (a) Sample images from the Weizmann-dataset [3]. (b) Some examples of different actions from different subjects and different cameras.](image-url)
is still a crucial problem. Several approaches to multi-view learning have been proposed in the machine learning literature [4, 12]. Multi-view learning exploits multiple redundant views to mutually train a set of classifiers defined in each view, which can be advantageous compared with learning with only a single view [4], especially when the weaknesses of one view complement the strengths of the other. Therefore, multi-view learning is an effective method to mine the effectiveness of different views.

On the other hand, it is difficult to model and measure different actions due to their diversity. Different cameras have different capture rate, therefore they produce the videos at different time resolutions. Even when using the same camera, different subjects and different environments also have an impact on the time resolution of action. What’s more, though they are different actions, they mainly contain similar shape features. Figure 1(b) illustrates an example. The video sequences (1) and (2) represent walk events, but they have different time resolutions due to different subjects or cameras. Though video sequences (1), (2), (3) and (4) represent three different actions, they contain some frames with similar shape feature. Thus, a time warping strategy should be performed and video matching should be conducted based on similar elements. To this end, the Gaussian Mixture Models (GMM) can be used to model action, and each Gaussian component is viewed as a video word. Therefore, we can transform length-variant orderless feature set of different action videos into a word frequency vector of a fixed length, and then conventional machine learning algorithms can be applied based on this fixed length representation.

In this paper, inspired by the idea of multi-view learning, we propose a novel framework (as showed in figure 2) to recognize action by extracting effective bag-of-words as a video clip representation using Co-EM strategy. Histograms of the silhouette and of the optical flow are extracted for each cropped frame centered on the human figure, and the feature vectors from all cropped frames within all training video clips are modeled using GMM. Then the Co-EM algorithm employs the two sets of features, i.e. two views, to solve the parameters of GMM. To automatically determine the number of GMM components, the Minimum Description Length (MDL) principle is adopted. Based on the learned GMM, each Gaussian component is viewed as a word, and each action video is described as a distribution of these words. Finally, we use an SVM as a classifier to train and test these actions.

2. RELATED WORK

The action representation and action modeling are two critical issues in human activity recognition. The features should be simple, intuitive and reliable to extract without manual labour. Some approaches exploit local descriptors based on interest points in images or videos. Schuld et al. [15] construct video representations in terms of local spatial-temporal features and integrated such representations with SVM for action recognition. Optical flow has also been widely used. Efros et al. [6] propose a spatio-temporal descriptor based on blurred optical flow measurements to recognize actions. The use of features available from silhouettes is increasingly popular. Blank et al. [3] utilize properties of the solution to the Poisson equation to extract features from space-time silhouettes for action recognition, detection and clustering. Daniel and Edmond [18] represent motion sequences with respect to a set of discriminative static key-pose exemplars and without modeling any temporal ordering. Du Tran and Alexander Sorokin [16] propose a metric learning based approach for human activity recognition. The feature they use is similar with ours, however we propose a multi-view learning based method using Co-EM strategy to syncretize the silhouette and optical flow features instead of simply splicing them.

To model and measure different actions, bag-of-words based approaches [11, 9] have been widely used to transform length-variant orderless feature set of action videos into a word frequency vector of a fixed length, and then a classifier is trained for action recognition based on this fixed length representation. The traditional methods to learn video words use k-means algorithm, which can not automatically decide the number of clusters. Jingen Liu and Mubarak Shah [9] propose an approach to automatically discover the number of video-word clusters by utilizing Maximization of Mutual Information (MMI). Based on our multi-view learning framework, we adopt the MDL principle to estimate the number of Gaussian components to obtain video words.

3. ACTION FEATURE EXTRACTION

3.1 Frame Description

The effectiveness of shape feature and optical flow feature has been demonstrated in action recognition [18, 2, 16]. Considering the two features can complement the strengths for each other, silhouette and optical flow are adopted to describe each frame of
video sequences. Silhouette can be obtained for instance with background subtraction, and the optical flow can be computed employing the Lucas and Kanade [10] algorithm. The optical flow vector field $F$ is then split into horizontal and vertical components of the flow, $F_x$ and $F_y$. To reduce the effect of noise, each of them is smoothed using median filter.

The input to our action recognition algorithm is a stabilized sequence of cropped frames, which are centered on the human figure. For each cropped frame, a template image with similar size is adopted to describe the silhouette and optical flow. The Fig. 3 shows an example of the template image, which is divided into many pie slices covering some degrees each. The maximum distance between the pixels in the cropped frame and the center of the cropped frame is calculated, and is quantized into $R_{bin}$ bins, which makes the description insensitive to changes in scale of the action. For the $r^{th}$ bin, each pie slice covers $\frac{\theta}{8}$ degrees and does not overlap. In our experiments, $\theta$ and $R_{bin}$ are set to be $30^\circ$ and 8 throughout our evaluation, respectively. The value of each bin is integrated over the domain of every slice. Because the cropped frame is not a circle, the histogram is only 417 dimensions, as showed in Fig. 3.

Each cropped frame is described with two histograms (silhouette and optical flow), which reflect global features of human action. Each bin in the histograms reflects local features. Experimental results demonstrate human actions are well characterized using this representation. To obtain compact description and efficient computation, the dimension of the features is further reduced using PCA.

### 3.2 Video Description

As we know, the video mismatch may exist in both spatial and temporal domains, that is, a cropped frame of one video clip may correspond to a cropped frame of another video clip belonging to the same action, but their positions and scales may be greatly different in both spatial and temporal domains. Therefore, video matching should be conducted based on smaller elements rather than whole frames or video clips. What’s more, it is necessary to consider the temporal ordering and time resolution of human action. To this end, we use a histogram of silhouette and of optical flow to describe each cropped frame and present the video as a bag of video words. The video words are the components of a global GMM, which is adopted to model different actions and describe the distribution of all histogram feature vectors. The reason to use a global GMM for characterizing the features of different actions is three-fold. First, the estimated GMM is a compact description of the underlying distribution of all histogram feature vectors, and is less prone to noise, compared with the histogram feature vectors themselves. Second, actions in the same category may have some cropped frames with different features and actions in the different category may have some cropped frames with similar features. Third, the Gaussian components in GMM impose an implicit multi-mode structure of the histogram feature vector distribution in a video clip.

The GMM is estimated using histogram feature vectors extracted from all training video clips, regardless of their action labels. Here we denote $h \in \mathbb{R}^D$ as a histogram feature vector, where $D$ is the feature dimension. The distribution of the variable $h$ is modeled by GMM as

$$p(h; \Theta) = \sum_{k=1}^{K} w_k N(h; u_k, \Sigma_k),$$

where $\Theta = \{w_1, u_1, \Sigma_1, \ldots, w_K, u_K, \Sigma_K\}$ are the weight, mean, and covariance matrix of the $k$th Gaussian component, respectively, and $K$ is the total number of Gaussian components. The density is a weighted linear combination of $K$ unimodal Gaussian densities, namely,

$$N(h; u_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(h-u_k)^T \Sigma_k^{-1} (h-u_k)}.$$

Then Co-EM strategy is proposed to obtain a maximum likelihood parameter set of the models instead of conventional EM. To automatically determine the number of video-word clusters, the MDL principle is adopted. These will be introduced in detail in Section 4.

### 4. LEARNING BAG OF VIDEO WORDS

Co-EM [12] is a semi-supervised, multi-view algorithm that uses the hypothesis learned in one view to probabilistically label the examples in the other one. Intuitively, Co-EM runs EM in each view and, before each new EM iteration, inter-changes the probabilistic labels generated in each view. Co-EM can be seen as a probabilistic version of Co-Training [4]. In fact, both algorithms are based on the same underlying idea: they use the knowledge acquired in one view (i.e., the probable labels of the examples) to train the other view. The major difference between the two algorithms is that Co-EM does not commit to a label for the unlabeled examples; instead, it uses probabilistic labels that may change from one iteration to the other. Compared with those only one-view based methods, the
methods based on the Co-EM algorithm have the following advantages [8]: compensate imperfectness of single view, be more reliability, improve local optimality and accelerate convergence rate. Inspired by the effectiveness of the multi-view learning, we use the well-known finite GMM to model the content of action, and propose the Co-EM strategy to estimate the parameters of GMM. The idea of Co-EM is using two classifiers (GMM) to be trained to find a suitable initialization for the GMM (Classifier1) that is subsequently adapted using Co-EM. Based on the initialization, the responsibilities $\rho_1$ are evaluated and transferred to the View2 to train the Classifier2 via EM in the $\text{View}_2$. Then the responsibilities $\rho_2$ in View2 are calculated and transferred to the View1. This process can be run iteratively until reaching the convergence of the log likelihood in the two views or the maximum number of iterations.

We have thus far not discussed how to choose $K$, the number of mixture components. A criterion was suggested by Rissanen [13] called the minimum description length (MDL) estimator. The objective is to minimize the MDL criterion given by

$$MLD(K, \Theta) = - \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} w_k N(h; u_k, \Sigma_k)\right) + \frac{1}{2} L \log(ND),$$

where $N$ is the number of data, and $L$ is the number of free parameters needed for a model with $K$ mixture components. In the case of a Gaussian mixture with full covariance matrices, we have $L = (K - 1) + KD + K \frac{D(2D+1)}{2}$. The MDL criteria considers the penalty term on the total number of data values $ND$. In practice, this is important since otherwise more data will tend to result in over fitting of the model. To obtain the number of $K$, we start with a large number of clusters, and then sequentially decrement the value of $K$. For each value of $K$, we will apply the Co-EM to update the parameters until we converge to a local minimum of the MDL functional. After we have done this for each value of $K$, we may simply select the value of $K$ and corresponding parameters that resulted in the smallest value of the MDL criteria. For details, please see [5].

5. EXPERIMENTAL RESULTS

We have tested our algorithm on two datasets: Weizmann human action dataset [3], KTH human motion dataset [15]. The default experiment settings are as follows. The histograms of optical flow and silhouette are considered as View1 and View2 respectively. We use SVM as the multi-classifier, and adopt the leave-one-out cross-validation strategy on KTH and Weizmann respectively.

5.1 Weizmann Dataset

The Weizmann dataset [3] (see Figure 4) contains 10 actions: bend (bend), jumping-jack (jack), jump-in-place (pjump), jump-forward (jump), run (run), gallop-sideways (side), jump-forward-one-leg (skip), walk (walk), wave one hand (wave1), wave two hands (wave2), performed by 9 actors. In these experiments, the optical flow of the corresponding cropped frames and the background-subtracted silhouettes which are provided with the Weizmann dataset are considered as two views respectively. 8 out of the 9 actors in the database are used to estimate the parameters of GMM using Co-EM strategy instead of conventional single view based EM, and the 9th is used for the evaluation. This is repeated for all 9 actors and the rates are averaged.

In Fig. 4, we show two example testing videos from each category with their corresponding video-words histograms on two views to demonstrate discrimination of the distribution of the learnt video-words clusters. Actions from the same category share the similar video-words clusters distribution. It is also clear to see from the peaks of these histograms that some video-words clusters are dominating in one action but not others. When specifically looking into the action of a person described in two views, one might note that they have the similar distribution of the learnt video-words clusters.

This demonstrates our Co-EM strategy has been convergent in different views. Table 1 shows our method with Co-EM strategy has achieved a better performance than conventional single view based EM. From this table, we can see our method has achieved a 100% accuracy with multi-view learning.

In comparison on the Weizmann dataset, the space-time volume approach proposed by Blank et al. [3] has a recognition rate of 99.61%, Wang and Suter [17] report a recognition rate of 97.78% with an approach that uses kernel-PCA for dimensional reduction and factorial conditional random fields to model motion dynamics. The work of Ali et al. [1] uses a motion representation based on chaotic invariants and reports 92.6%. Daniel and Edmond [18] report a recognition rate of 100% using exemplars. Alireza Fathi and Greg Mori [7] construct mid-level motion features which are built from low-level optical flow information for action recognition, they report a recognition rate of 100%. Note, however, that a precise comparison between the approaches is difficult, since experimental setups, e.g. number of actions and length of segments, slightly differ with each approach. Based on the Co-EM strategy, the strengths of the different views, silhouette and optical flow, are complemented for each other and the accuracy has been improved for each view. This shows our multi-view learning framework is effective and precise.

5.2 KTH Dataset

The KTH human motion dataset (see Fig 5(a)) contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping). Each action is performed several times by 25 subjects in four different conditions: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In this experiment, we used edge filtered sequences instead of background subtracted silhouettes. Edges are detected independently using a Canny edge detector and optical flows are calculated in each frame of the original sequences. Based on the locations of people detected by using the method in Sahmeydani and Mori [14], the histograms of edge and optical flow for each cropped frame can be extracted.
PCA.

ing with \{...\} to directly applying k-means. We perform twenty different clusterings using Co-EM algorithm and

Table 1: The comparison of different methods on Weizmann dataset. The dimension of feature in View1 and View2 are 60 using PCA.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Action</th>
<th>View1 (%)</th>
<th>View2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bend</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Jack</td>
<td>88.89</td>
<td>77.78</td>
</tr>
<tr>
<td>EM View1</td>
<td>Jump</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>EM View2</td>
<td>Pjump</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Co-EM View1</td>
<td>Run</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Co-EM View2</td>
<td>Side</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Skip</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Walk</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Wave1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Wave2</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2: The comparison of different methods about mean accuracy on KTH dataset. STIP means spatio-temporal interest points.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean Accuracy</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schultet et al [15]</td>
<td>91.71%</td>
<td>STIP</td>
</tr>
<tr>
<td>Niebles and Fei-Fei [11]</td>
<td>95.50%</td>
<td>STIP</td>
</tr>
<tr>
<td>Saad Ali and Munir Shah [1]</td>
<td>87.61%</td>
<td>Optical Flow</td>
</tr>
<tr>
<td>Jingen Liu and Mubarak Shah [9]</td>
<td>94.10%</td>
<td>3D interest points</td>
</tr>
<tr>
<td>Our method</td>
<td>95.33%</td>
<td>Edge + Optical Flow</td>
</tr>
</tbody>
</table>

Because this dataset contains tens of thousands of cropped frames, we first uniformly subsample the sequences by a factor of 1/20 in training dataset and then use Co-EM strategy to estimate the parameters of GMM. We use 24 videos of actors as training dataset and the rest as testing videos, and the results are reported as the average accuracy of 25 runs. The confusion matrix for this experiment is shown in Fig 5(c). The number of video-words clusters is \(K = 50\), and the average accuracy is 95.33\%. From these tables, we can see “jogging” and “running” are confused, because the two actions are very similar in the two views.

Here, we also investigate the gain of Co-EM algorithm compared to directly applying k-means. We perform twenty different clustering with \{5, 10, \cdots , 95, 100\} clusters using Co-EM algorithm and k-means algorithm respectively. Fig 5(b) shows the results. From the figure, we can see that Co-EM algorithm can significantly improve the performance in View2 compared with k-means, however, the performance in View1 is not improved distinctly. The reason is the feature in View2 extracted by Canny edge detector is not robust and is very weak, therefore, it does not complement the strength of View1. What is more, the k-means in View1 having a good performance demonstrates the feature extracted is effective.

We also compare our performance with other state-of-art algorithm on KTH dataset. The performance is reported in Table 2. It can be seen that performance using our proposed multi-view learning framework exceeds other methods. We believe the reason for the improvement is that our method complements the strengths of different views by Co-EM algorithm, in addition, the GMM is efficient and robust to model different actions.

6. CONCLUSION

In this paper, we proposed a novel framework to recognize action by extracting efficient bag-of-words as a video clip representation. The static visual information and motion information were considered as two views (features) to represent each frame and GMM was adopted to model the distributions of those features. To automati-
(a) Example actions from the KTH data set.

(b) The comparison of k-means and Co-EM in two views, respectively.

(c) Confusion matrix for action recognition

Figure 5: The experimental results on KTH dataset

cally determine the number of GMM components, the MDL principle was used. Instead of conventional single view based EM, the Co-EM strategy was proposed to estimate the parameters of GMM. This strategy complemented the strengths of the different features (views) and improved the performance action recognition. Our approach had been extensively tested on two public datasets: KTH and Weizmann datasets, and we obtained some good performance on both dataset. To the best of our knowledge, we were the first to apply the Co-EM based multi-view learning approach on action recognition, and had obtained competitive results. For the future work, We will extend the method for other features, such as spatio-temporal interest points, and recognizing realistic human actions in unconstrained videos such as in feature films.

7. ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (Project No. 60835002 and 90920303), and National Basic Research Program (973) of China (Project No. 2010CB327905).

8. REFERENCES