Spatiotemporal Pyramid Representation for Recognition of Facial Expressions and Hand Gestures

Zhipeng Zhao , Ahmed Elgammal
Computer Science Department, Rutgers University
110 Frelinghuysen Road, Piscataway, NJ 08854-8019, U.S.A
{zhipeng,elgammal}@cs.rutgers.edu

Abstract

This paper presents a spatiotemporal pyramid representation for recognizing facial expressions and hand gestures. This approach works by partitioning video sequence into increasingly fine subdivisions in the space and time domains and modeling the distribution of the local motion features inside each subdivision such that the set of motion features are mapped into spatial and temporal multi-resolution histograms. This spatiotemporal pyramid is built by weighting the histograms from the different layers of the subdivisions. The proposed approach is an extension of the orderless "bag-of-words" model by approximately capturing geometric and temporal arrangements of the local motion features. The experiments on facial expression and hand gesture data sets have demonstrated the significantly improved performance over state of art results on human activity recognition tasks by using our representation.

1. Introduction

Recognizing human activities from image sequences is an appealing yet challenging problem in computer vision. Reliable and effective solutions to this problem can be applied to many areas including motion capture, human-computer interaction, environment control, and security surveillance. In this paper, we focus on recognizing the activities of a person, e.g. facial expressions and hand gestures, in an image sequence from local motion features and their spatiotemporal arrangements.

Our approach is motivated by the recent success of "bag-of-words" model for general object recognition in computer vision[19, 13]. This representation, which is adapted from text retrieval literature, models the object by the distribution of words from a fixed visual code book, which is usually obtained by vector quantization of local image visual features. However, this method discards the spatial and temporal relations among the visual features, which could be helpful in the object recognition. Addressing this problem, our approach uses a hierarchical structure in the video sequence representation to integrate information from the spatial and temporal domains. We first apply a spatiotemporal feature detector to the video sequence and obtain the local motion features. Then we generate a visual word code book by quantization of the local motion features and assign word label to each of them. Next we divide the data volume spatially and temporally into finer subdivisions and compute the histograms of the visual words in each cell. Finally, we concatenate the histograms from all cells and use it as the feature for the whole video sequence. The contribution of our work lies in that besides the appearance information contained in the local motion features, our representation also captures both the spatial and the temporal relation among these motion features, which leads to better performance than the popular "bag-of-words" approach.

The organization of this paper is as follows. The related work are summaries in Section 2. In Section 3 we introduce the framework and in Section 4 we applied this framework as the spatiotemporal pyramid representation built from motion features described in Section 5. Section 6 presents the results of applying the proposed representation on facial expression and hand gesture data sets. Section 7 is the discussion.

2. Related Work

Extensive research has been done in recognizing human activities. The approaches can be broadly categorized as model based, spatiotemporal template based and "bag-of-words" based. Model based approaches for activity recognition depend on locating and tracking body limbs in order to recognize the activity. That requires a model of the body, whether a 3D model or a 2D view-based model. We refer the reader to excellent surveys covering this topic, such as [2, 6, 11]. However, for the task of activity recognition, tracking the limbs is not necessary. That motivates research on obtaining spatiotemporal descriptors directly from the

978-1-4244-2154-1/08/$25.00 2008 IE

motion to recognize the activity without limb tracking. One of the earliest work on spatiotemporal descriptor was carried out by Polana and Nelson[14]. In Bobick and Davis's work[3], Motion-Energy-Image and Motion-History-Image are introduced as templates for different motion recognition. Efros et al. [5] also proposed a spatiotemporal descriptor based on global optical flow measurements. Spatiotemporal template approaches are holistic approaches where global descriptors are used with no local features extracted.

In contrast, "bag-of-words" based approaches detect local salient descriptors as visual words, which are then used to recognize the activity. "bag-of-words" has been used successfully for object categorization[19, 13]. Inspired by text categorization, it represents the object as histogram of local features. Recently, "bag-of-words" methods have been used in activity recognition[16, 4, 17]. However, "bag-of-words" approaches solely depend on the dense local motion features. They lack the relations between the features in the spatial and the temporal domains. These are important correlation, which are helpful for recognition. There are many recent research on extending "bag-of-words" to add the spatial relation in the context of object categorization [15, 1, 10, 7, 9]. In particular, pyramid match kernel [7] used the weighted multi-resolution histogram intersection as a kernel function for classification with sets of image features. Also, spatial pyramid approach[9] introduced a spatial pyramid structure to find approximate correspondence at different levels between two sets.

The approach we propose here tries to simultaneously capture the spatial and temporal distribution of the local motion features by building a hierarchical structure along both spatial and temporal dimensions for representation. This approach is inspired by [7, 9]. However, our approach is a representation which embeds the spatial and temporal information while [7, 9] proposed a matching kernel. Our approach uses Chi-square distance as a distance function while [7, 9] used weighted histogram intersection to satisfy as a kernel function.

Other related directions and extensions for "bag-of-words" in the context of activity recognition include [8, 18, 20]. In [8]'s work, spatial orientation information were captured in the local features. In [18, 20], latent semantic model was applied to discover the activity types as topics in the hidden layer between the visual features and the video sequence. Different from them, our approach uses the spatiotemporal pyramid as a representation and simultaneously integrates the spatiotemporal relation among visual features with their appearance information.

3. Spatial Temporal Pyramid Representation

3.1. Pyramid representation for sets of points

The goal of our approach is to find a representation for a set of points in a spatiotemporal space. This representation should capture the distribution of a set of points in the space in a way that it is suitable for measuring the similarity between the sets. Inspired by [7, 9], we propose a pyramid representation to capture the spatiotemporal distribution for sets of points.

Let X and Y be two sets of points in a d-dimensional space, $X = \{x_i | x_i \in R^d\}$, $Y = \{y_j | y_j \in R^d\}$. We recursively partition the space into subdivisions. Intuitively, we measure the distance between X and Y as the sum of the distances between the distributions of the points from each data set in each of the subdivisions. The distributions of the points are approximated by the number of points in the subdivision whose distances are measured by Chi-square distance.

We start by constructing a sequence of increasingly finer binary partitions at resolution 0,..., L, such that the partition at level l has 2^l cells along each dimension with a total of $D = 2^{dl}$ cells for this level. We denote the number of points from X in the ith cell at level l as $H_X^l(i)$. The distance between X and Y at this level, represented as H_X^l, H_Y^l respectively, is the sum of the distances for each corresponding cell,i.e.:

$$dist(H_X^l, H_Y^l) = \sum_{i=1}^{D} \chi^2(H_X^l(i), H_Y^l(i)) \qquad (1)$$

where $\chi^2(\cdot,\cdot)$ is the Chi-square distance. This is similar to representing the sets of points X and Y at level l by concatenating $H_X^l(i)$ and $H_Y^l(i)$ into histograms respectively and measuring their Chi-square distance. Therefore, we can use these concatenated histograms as the representations for H_X^l and H_Y^l respectively.

In this pyramid structure, different level captures different scale of variance. The representation for a set of points X should include all $H_X^l, l = 0, .., L$ and the distance between X and Y, represented as H_X and H_Y respectively, should include the distances from all levels:

$$dist(H_X, H_Y) = \sum_{l=0}^{L} dist(H_X^l, H_Y^l) \qquad (2)$$

This is again similar to representing X and Y by concatenating their histogram representations from all levels into a long histogram respectively and measuring their distance. Therefore, we can use these concatenated histograms as the representations for H_X and H_Y respectively.

3.2. Weighed pyramid representation for sets of points

Since different information are captured at various levels of the pyramid, different weights should be assigned for each of them. At finer resolution, the correspondence between two sets is captured more accurately. Therefore, we penalize the similarity information gained at a coarser level

and give more weights to the similarity measured by the histogram distance at a finer resolution. The weight we assign at level l is: $weight(l) = 1/2^{L-l}$ for $l = 0, .., L$. The weighted distance between X and Y is:

$$dist(H_X, H_Y) = \sum_{i=0}^{L} \frac{1}{2^{L-l}} dist(H_X^l, H_Y^l) \quad (3)$$

Since Chi-square distance satisfies

$$c\chi^2(a, b) = \chi^2(ca, cb) \quad (4)$$

where c is a scalar, we can directly embed the weight to the histogram representation. Putting everything together, our representation for a set of points X is the concatenated weighted histogram from all levels of the pyramid.

Since the distance between these representations is the sum of the Chi-square distances between each element, which by themselves are metric, it is easy to prove that:

1) $dist(H_X, H_X) = 0$

2) $dist(H_X, H_Y) = dist(H_Y, H_X)$

3) $dist(H_X, H_Z) \leq dist(H_X, H_Y) + dist(H_Y, H_Z)$

Therefore, the distance defined on our representation is a metric.

Our representation is suitable for the set of points in a spatiotemporal space because it is an approximation of the distribution of the points in the spatiotemporal space. Since it is histogram based representation, Chi-square distance is suitable for measuring the similarity between the sets of points.

In Lazebnik et al.'s work[9], pyramid match kernels are proposed to use a pyramid structure to find approximate correspondence at different levels between two sets. Our work differs from them in that:

1) Our goal is to find a suitable representation to integrate the spatial and temporal relation for a set of points. It embeds the weights in the representation to reflect the importance of different pyramid layers. The work in [9] is seeking a suitable kernel function for two sets of points.

2) Because our representation is a concatenated histogram, we measure the distance by Chi-square distance. The pyramid match kernels use histogram intersection as the distance function to satisfy the Mercer's condition.

3) Our representation captures the distribution of the points in both spatial and temporal space, so it can be applied for human action recognition in the spatiotemporal domain. The pyramid matching kernels are only used for natural scene categorization in 2-D images.

4. Pyramid representation for human action

Motivated by the "bag-of-words" approach while still considering the spatial and temporal arrangements of the features, we model human activity as a set of local motion features points located in the three dimensional spatiotemporal space. Then the spatiotemporal pyramid representation can be used for the set of feature points. We divide the video sequence spatially and temporally into increasingly finer subdivisions and compute the distribution of the feature points in each cell for all levels. The final representation for the activity is the concatenated weighted histogram from all levels.

We also want to consider the appearance information of the local motion features and model them as words. We apply k-means clustering in the visual feature space to quantize all local motion features into K discrete types and assign word label to each of them. In each subdivision, we use the histogram of the visual words instead of the number of points as the approximation for the point distribution. This representation contains both appearance information, as the histogram in each cell, and the spatiotemporal information, which comes from the spatiotemporal pyramid structure.

This representation is a straightforward extension of the popular "bag-of-words" method. In each subdivision, all the local motion features are modeled as "bag-of-words". When $L = 0$, it reduces to the standard "bag-of-words" representation. For better computation efficiency, we normalize the vector by the total weights of all elements.

The complexity of this representation is linear with the size of motion words vocabulary. For L level and K motion words, the dimensionality of the resulting representation is $K \sum_{l=0}^{L} 8^l = K\frac{1}{7}(8^{L+1} - 1)$. In our experiments we observe that the performance does not improve much when $L > 2$. Therefore, we use the setting of $K = 250$ and $L = 2$, which leads to a 18250-dimension vector for the activity representation.

5. Feature Extraction

This section briefly describes the local motion features used in the experiments of Section 6. There are various methods for motion feature detection and representation, such as presented in [16] and [4]. As noticed by [4] and observed from our experiments, the interest points detected by generalized space-time interest points detector from [16] are too sparse to build model for many complex activities. Therefore, we utilized the one from Dollar[4], which has been proven successful in [4, 12, 20]. Here we give a brief review of this method.

Like many interest point detectors, in [4], the space-time interest points are detected by applying separable linear filter to the video sequences. With the assumption of a stationary camera or a preprocess to account for the camera motion, the response function has the following form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (5)$$

where $g(x, y; \sigma)$ is the $2D$ Gaussian smoothing kernel applying along the spatial dimension (x, y), with parame-

Dataset	Facial Expression	Hand Gesture
No. of classes	6	9
No. of subjects	2	2
No. of trials per subject	8	10
No. of capturing condition	2	5
Total No. of Samples	192	900

Table 1. Details of the data sets used in our experiments.

ter σ corresponding to the spatial scale of the detector. h_{ev} and h_{od} are a quadrature pair of $1D$ Gabor filter applying along the temporal dimension. They are defined as $h_{ev}(t;\tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2}$, with parameter τ corresponding to the temporal scale of the detector. In all cases, we chose $\omega = 4/\tau$, as did in [4].

To represent the motion feature, a cuboid of spatiotemporally windowed data surrounding the detected interest point (local maxima of response function) is extracted. In our experiments, it is set to be six times the scale of the detector to contain the volume contributing to the response function. We then compute the gradients of the intensities in the cuboid and flatten them into a vector. Finally, we project the vectors into a low dimensional space by PCA and use the more compact representations as the motion features for the video sequences.

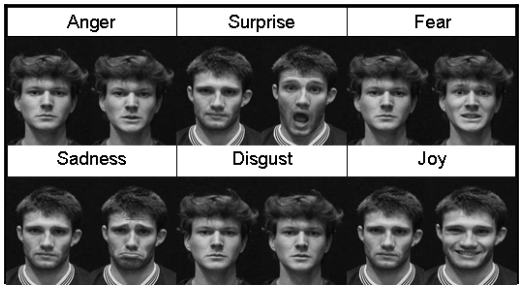
To build the code book, we perform k-means from a random subset of motion features from the training data. The typical vocabulary size for our experiments is K=250.

6. Experiments

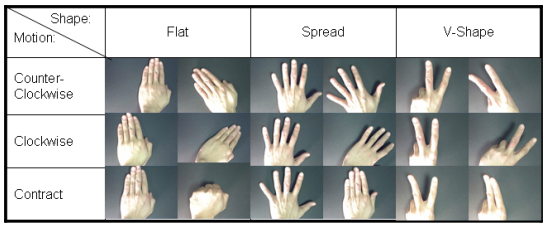
6.1. Data sets

We carried out our experiments in two applications, namely facial expressions and hand gestures recognition. We used the facial expressions data set, which was introduced by Dollar et al.[4] and the hand gestures data set, which was collected by Wong et al.[20]. In both data sets, each video sequence contains one activity, starting and ending from the neutral position. The video sequences were converted into gray level to avoid the bias in color. The details of the data sets are summarized in Table 1 and some sample images from the video sequences are shown in Figure 1.

For both data sets, we carried out a baseline approach using the "bag-of-words" representation. We also collected all the published experiment results we could find on these two data sets. To better compare with previous studies, we used the same nearest neighbor classification algorithm for recognition. Since there were some randomness factors in building the code book by $k-$means clustering method, all results were obtained by averaging over 20 runs.



(a) Facial expression data set



(b) Hand gesture data set

Figure 1. Sample images from the experiment data sets.

6.2. Facial Expression

With the same experiment setting as in [4], we trained on one subject under one of the two lighting conditions and tested on: (1) the same subject under the same illumination, (2) the same subject under different illumination, (3) a different subject under the same illumination, and (4) a different subject under different illumination. Since Dollar's implementation[4] used a "bag of words" approach, we used it as the baseline algorithm. We compared the confusion matrices in the first two scenarios from the two approaches in Figure 2. In the same subject under the same illumination scenario, the recognition task was easy. The baseline algorithm already achieved very hight recognition rate, so our approach only slightly improved the results. In the same subject under different illumination scenario, our approach has shown great improvements for all facial expression types.

We also show in Table 2 the recognition rates in all four different scenarios from the baseline "bag-of-words" algorithm and spatiotemporal pyramid representation with different number of layers. From the baseline algorithm, the average recognition rate across different scenarios is 73.92%. Both recognition rates from our pyramid representation with different number of levels have shown significantly improved performance. And the best result is 86.38% from the 2-level configuration.

We also tested on the facial expressions data set with the

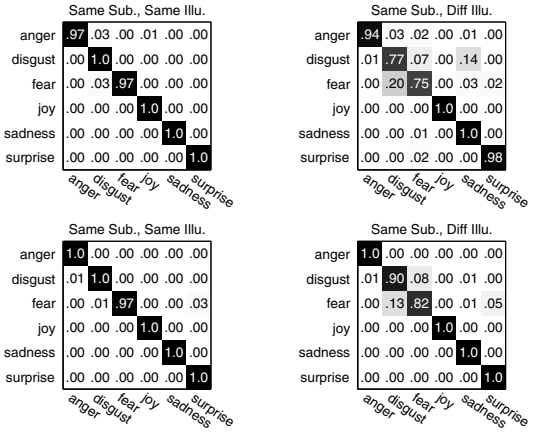


Figure 2. Comparison of recognition rates in the first two scenarios. The top row shows the confusion matrices from Dollar's implementation[4] and the bottom row shows the confusion matrices from our 1 level spatiotemporal pyramid representation.

Methods	Same Sub. Same Illu.	Same Sub. Diff. Illu.	Diff Sub. Same Illu.	Diff Sub. Diff Illu.
Baseline	98.83	90.46	58.67	47.71
1 level Pyramid	**99.33**	**95.29**	78.33	69.29
2 level Pyramid	98.17	94.75	**78.92**	**73.67**

Table 2. The facial expression recognition rates(%)in different scenarios from the baseline "bag-of-words" algorithm and the spatiotemporal pyramids with different number of levels.

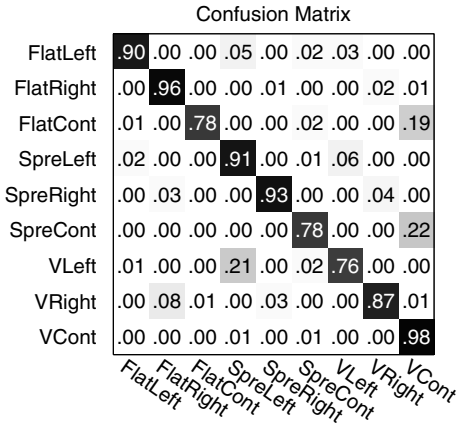
Methods	Accuracy (%)
Base line	91.33
Pyramid (1 level)	**94.67**
Pyramid (2 level)	93.50
pLSA [20]	50.00
pLSA-ISM [20]	83.33

Table 3. The facial expression recognition rates obtained from different algorithms with leave-one-out cross-validation experiment setting.

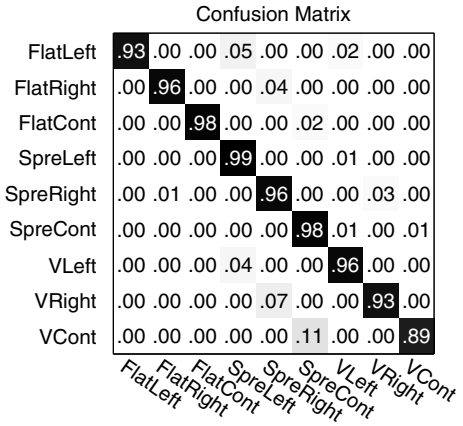
same experiment setting as in [20], which was the leave-one-out cross-validation. The average recognition rates from the different algorithms for the six types of facial expressions are shown in Table 3. This has shown the pyramid representation can improve on the "bag-of-words" baseline model and even achieve better performance than the complicated probabilistic latent semantic models[20].

6.3. Hand Gesture

We used the leave-one-out cross-validation experiment setting, which is the same as in [20], for hand gesture recognition. In this experiment, the video from one objects under one capturing condition was used in testing and the remaining was used in training.



(a) Confusion Matrix using pLSA-ISM[20]



(b) Confusion Matrix using 2 level spatial-temporal pyramid

Figure 3. The confusion matrices for recognizing the nine types of hand gestures. The top confusion matrix is obtained by using pLSA-ISM[20] and the bottom confusion matrix is from our work by using 2 level spatiotemporal pyramid representation.

The confusion matrices for recognizing the nine types of hand gestures are shown in Figure 3. The top confusion matrix is obtained by using pLSA-ISM[20] and the bottom confusion matrix is obtained from our approach by using 2 level spatiotemporal pyramid representation. It shows that by using the pyramid representation, the recognition rate has improved on every categories except the last one.

The average recognition rates from different algorithms for all hand gestures are shown in Table 4. This has shown the pyramid representation can achieve much better recognition rate than the "bag-of-words" approach and even exceed the results from complicated probabilistic latent semantic models[20].

From experiments on both data sets, we do not observe any significant increase in performance beyond 2-level pyramid configuration. This is because when $l = 2$,

Methods	Accuracy (%)
Base line	85.81
Pyramid (1 level)	95.31
Pyramid (2 level)	**96.44**
pLSA [20]	76.94
pLSA-ISM [20]	91.94

Table 4. The hand gesture recognition rates obtained from different algorithms.

the 64 subdivisions of the whole video sequence already roughly capture the sets of points' locations in the spatiotemporal domain while maintain tolerance for the locations variance in each cell. With more levels, the number of features points falling into each cell will be decreased, so the histograms might not be a good approximation for the feature distribution.

7. Discussion

We have presented a spatiotemporal pyramid representation for human activity recognition. Our approach simultaneously integrates the spatiotemporal relation among local motion features with their appearance information and embeds these rich information in the pyramid representation for the video sequence. Currently we are working on video sequences which contain activity starting and ending in a neutral position without global motion. This makes it easy to partition the sequence in the spatial and temporal domains. In the future, we intend to investigate methods to detect periodicity of the activity and compensate for global motion such that our approach can be applied in more general scenario.

Acknowledgments

This research is partially funded by NSF CAREER award IIS-0546372.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures: Multilevel local coding for visual recognition. In *ECCV06*, pages I: 30–43, 2006. 2

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999. 1

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 2

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005. 2, 3, 4, 5

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 726, 2003. 2

[6] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999. 1

[7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1458–1465, 2005. 2

[8] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *HUMO07*, pages 271–284, 2007. 2

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006. 2, 3

[10] M. Marszaek and C. Schmid. Spatial weighting for bag-of-features. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2118–2125, 2006. 2

[11] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, 2006. 1

[12] J. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC06*, page III:1249, 2006. 3

[13] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006. 1, 2

[14] R. Polana and R. Nelson. Detecting activities. In *DARPA93*, pages 569–574, 1993. 2

[15] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2033–2040, 2006. 2

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR04*, pages III: 32–36, 2004. 2, 3

[17] C. Thurau. Behavior histograms for action recognition and human detection. In *HUMO07*, pages 299–312, 2007. 2

[18] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *HUMO07*, pages 240–254, 2007. 2

[19] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorization nine visual classes using local appearance descriptors. In *IWLAVS*, 2004. 1, 2

[20] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR07*, pages 1–6, 2007. 2, 3, 4, 5, 6