Gesture Recognition Under Small Sample Size

Tae-Kyun Kim[1] and Roberto Cipolla[2]

[1] Sidney Sussex College, University of Cambridge, Cambridge, CB2 3HU, UK
[2] Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK

Abstract. This paper addresses gesture recognition under small sample size, where direct use of traditional classifiers is difficult due to high dimensionality of input space. We propose a pairwise feature extraction method of video volumes for classification. The method of Canonical Correlation Analysis is combined with the discriminant functions and Scale-Invariant-Feature-Transform (SIFT) for the discriminative spatiotemporal features for robust gesture recognition. The proposed method is practically favorable as it works well with a small amount of training samples, involves few parameters, and is computationally efficient. In the experiments using 900 videos of 9 hand gesture classes, the proposed method notably outperformed the classifiers such as Support Vector Machine/Relevance Vector Machine, achieving 85% accuracy.

1 Introduction

Gesture Recognition Review
Gesture recognition is an important topic in computer vision because of its wide ranges of applications such as human-computer interfaces, sign language interpretation and visual surveillance. Not only spatial variation but also temporal variation among gesture samples make this recognition problem difficult. For instance, different subjects have different hand appearance and may sign gesture in different pace.

Recent work in this area tends to handle the above variations separately and therefore leads to two smaller areas, namely posture recognition (static) and hand motion or action recognition (dynamic). In posture recognition, the pose or the configuration of hands is recognised using silhouette [5] and texture [6]. By contrast, hand motion or action recognition interprets the meaning of the movement using full trajectory [9], optical flow [4] and motion gradient [11].

Compared with hand motion recognition, posture recognition is easier in the sense that state-of-the-art classifiers, e.g. Support Vector Machine, Relevance Vector Machine [11] or Adaboot [6] can be directly applied to it. Gesture recognition, on the other hand, has adopted rather different approaches, e.g. Hidden Markov Model [9]) or Dynamic Time Warping [3]), to discriminate dynamic/or temporal information which is typically highly non-linear in a data space. These methods, especially the Hidden Markov Models, have many parameters to set, a large amount of training examples, and difficulty for extension to large vocabulary [2]. Besides, these traditional methods have not integrated the posture and

Y. Yagi et al. (Eds.): ACCV 2007, Part I, LNCS 4843, pp. 335–344, 2007.

temporal information and thus are difficult to differentiate gestures of similar movements signed by different hand shapes.

Some recent works [8] directly operate with full spatiotemporal volume considering both posture and temporal information of gestures to a certain degree, but are still unreliable in cases of motion discontinuities and motion aliasing. Also, the method [8] requires the manual setting of the important parameters such as positions and scales of local space-time patches. Another important line of methods exploits visual code words (for representation) with either a Support Vector Machine (SVM) or a probabilistic generative model [12,13]. Again, for their good performance, it is critical to properly set the parameters associated with the representation, for e.g. space-time interest points and code book size.

Motivation and Summary of This Study

To avoid empirical setting of the parameters in the existing methods, it seems obvious to seek a more generic and simpler learnable approach for gesture recognition. Note that many of critical parameters in the previous methods are incurred in the step of representing gesture videos prior to using classifiers. In that case, it could be better to apply learnable classifiers directly to the videos which can be simply converted into column vectors. Unfortunately, this is not a good way either. Vectorization of a video by concatenating all pixels in the three-dimensional video volume causes a high dimension of N^3, which is much larger than N^2 of an image. Also, it may be more difficult to collect sufficient number of video samples for classifiers than images (see that a single video consists of multiple images). So called small sample size problem is more serious in learning classifiers with videos than images.

Getting back to the representation issue, this work focuses on how to learn useful features from videos for classification, discussing its benefits over direct using classifiers. With the given discriminative features, even a simple Nearest Neighbor classifier (NN) achieved a very good accuracy. An extension of Canonical Correlation Analysis (CCA) [1,15]-a standard tool of inspecting linear relationships of two sets of vectors- is proposed to yield robust pairwise features of any two gesture videos. The proposed method is closely related to our previous framework of Tensor Canonical Correlation Analysis [14], which extends the classical CCA into multidimensional data arrays by sharing either a single axis or two axes. The method of sharing two axes, i.e. *planes* between two video data, is updated and combined with the discriminative functions and the Scale-Invariant-Feature-Transform for further improvements. The proposed method does not require any significant meta-parameters to be adjusted and can learn both posture and temporal information for gesture classification.

The rest of the paper is organized as follows: Next section explains the proposed method with the discriminant functions, discussing the benefit of the method over traditional classifiers. The SIFT representation for video data is combined to the method for improvements in Section 3. Section 4 shows the experimental results and Section 5 draws conclusion.

2 Discriminative Spatiotemporal Canonical Correlations

Canonical Correlation Analysis (CCA) has been a standard tool of inspecting linear relationships of two random variables, or two sets of vectors. This was recently extended to two multidimensional data arrays in [14]. The method of spatiotemporal canonical correlations (which is related to the previous work in exploiting *planes* rather than *scan vectors* of two videos) is explained as follows: A gesture video is represented by firstly decomposing an input video clip (i.e. a spatiotemporal volume) into three sets of orthogonal planes, namely XY-, YT- and XT-planes as shown in Figure 1. This allows posture information in XY-planes and joint posture/dynamic information in YT and XT-planes. Three kinds of subspaces are learnt from the three sets of planes (which are converted into vectors by raster-scanning). Then, gesture recognition is done by comparing these subspaces with the corresponding subspaces from the models by classical canonical correlation analysis, which measures *principal angles* between subspaces[1]. By comparing subspaces of an input and a model, robust gesture recognition can be achieved up to pattern variations on the subspaces. The similarity of any model $\mathcal{D}_m$ and query spatiotemporal data $\mathcal{D}_q$ is defined as the weighted sum of the normalized canonical correlations of the three subspaces by

$$\mathcal{F}(\mathcal{D}_m, \mathcal{D}_q) = \Sigma_{k=1}^{3} w^k \mathcal{N}^k(\mathbf{P}_m^k, \mathbf{P}_q^k) \tag{2}$$

where,

$$\mathcal{N}^k(\mathbf{P}_m^k, \mathbf{P}_q^k) = (\mathcal{G}(\mathbf{P}_m^k, \mathbf{P}_q^k) - m^k)/\sigma^k, \tag{3}$$

$\mathbf{P}^1, \mathbf{P}^2, \mathbf{P}^3$ denotes a matrix containing the first few eigenvectors in its columns of XY-planes, XT-planes, YT-planes respectively and $\mathcal{G}(\mathbf{P}_m, \mathbf{P}_q)$ sum of the canonical correlations computed from $\mathbf{P}_m, \mathbf{P}_q$. The normalization parameters with index k are mean and standard deviation of matching scores, i.e. $\mathcal{G}$ of all pairwise videos in a validation set for the corresponding planes.

The discriminative spatiotemporal canonical correlation is defined by applying the discriminative transformation [10] learnt from each of the three data domains as

$$\mathcal{H}(\mathcal{D}_m, \mathcal{D}_q) = \Sigma_{k=1}^{3} w^k \mathcal{N}^k(h(\mathbf{Q}^{kT}\mathbf{P}_m^k), h(\mathbf{Q}^{kT}\mathbf{P}_q^k)), \tag{4}$$

[1] Canonical correlations between two d-dimensional linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$ are uniquely defined as the maximal correlations between any two vectors of the subspaces [1]:

$$\rho_i = \cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{L}_1} \max_{\mathbf{v}_i \in \mathcal{L}_2} \mathbf{u}_i^T \mathbf{v}_i \tag{1}$$

subject to: $\mathbf{u}_i^T \mathbf{u}_i = \mathbf{v}_i^T \mathbf{v}_i = 1,\ \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0,\ j = 1, ..., i-1$. We will refer to $\mathbf{u}_i$ and $\mathbf{v}_i$ as the i-th pair of canonical vectors. Multiple canonical correlations are defined by having next pairs of canonical vectors orthogonal to previous ones. The solution is given by SVD of $\mathbf{P}_1^T \mathbf{P}_2$ as

$$\mathbf{P}_1^T \mathbf{P}_2 = \mathbf{L} \Lambda \mathbf{R}^T \quad \text{where} \quad \Lambda = \text{diag}\{\rho_1, ..., \rho_d\}.,$$

where $\mathbf{P}_1, \mathbf{P}_2$ are the eigen-basis matrix, $\mathbf{L}, \Lambda, \mathbf{R}$ are the outputs of SVD.

where h is a vector orthonormalization function and $\mathbf{Q}^k$ are the discriminative transformation matrix learnt over the corresponding sets of planes. The discriminative matrix is found to maximize the canonical correlations of within-class sets and minimizes the canonical correlations of between-class sets by analogy to the optimization concept of Linear Discriminant Analysis (LDA) (See [10] for details). On the transformed space, gesture video classes are more discriminative in terms of canonical correlations. In this paper, this concept has been validated not only for the spatial domain (XY-subspaces) but also for the spatiotemporal domains (XT-, YT-subspaces).

Discussions

The proposed method is a namely *divide-and-conquer* approach by partitioning original input space into the three different data domains, learning the canonical correlations on each domain, and then aggregating them with proper weights. By this way, the original data dimension N^3, where N is the size of each axis, is reduced into $3 \times N^2$ so that the data is conveniently modelled. As shown in Figure 2a-c, each data domain is well-characterized by the corresponding low-dimensional subspace (e.g. hand shapes in XY-planes, joint spatial and temporal information in YT-, and XT- planes).

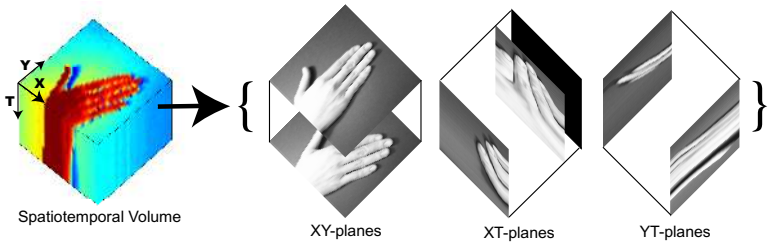


Fig. 1. Spatiotemporal Data Representation

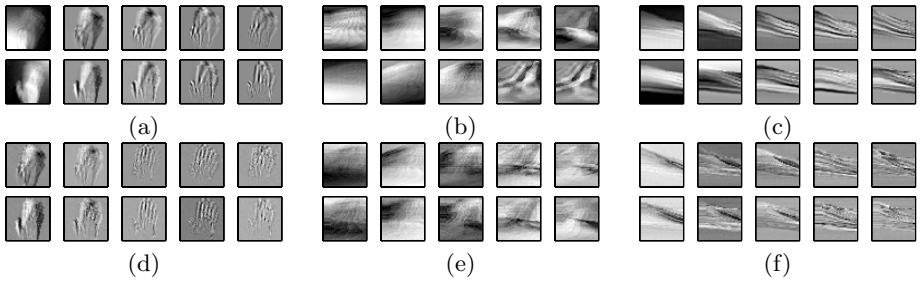


Fig. 2. Principal Components and Canonical Vectors: The first few principal components of the (a) XY (b) XT (c) YT subspaces of the two different illumination sequences of the same gesture class (See Figure 5) are shown at the top and bottom row respectively. The corresponding pairwise canonical vectors are visualized in (d) - (f). Despite the different lighting conditions of the two input sequences, the canonical vectors in the pair (top and bottom) are very much alike, capturing common modes.

Moreover, the method is robust using mutual (or canonically correlated) components of the pairwise subspaces. By finding the mutual components of maximum correlations, which are canonical correlations, some undesirable information for classification can be filtered out. See Figure 2 for the principal components and canonical vectors for the given two sequences of the same gesture class which were captured under the different lighting conditions. Whereas the first few principal components mainly corresponded to the different lighting conditions (in Figure 2a-c), the canonical vectors (in Figure 2d-f) well captured the common modes of the two sequences, being visually same in each pair. In other words, the lighting variations across the two sets were removed in the process of CCA, as it is invariant to any variations on the subspaces. Many previous studies have told that lighting variations are often confined to a low-dimensional subspace.

In summary, the proposed method has a benefit over direct learning classifiers under small sample size as drawn in Figure 3. High dimensional input space and a small training set often cause overfitting of classifiers to the training data and poor generalization to new test data. Distribution of the test samples taken under different conditions can be largely deviated from that of the training set, resulting in the majority of the test samples of class 1 misclassified in Figure 3. Nevertheless, the two intersection sets of the train and test sets are still placed in the correct decision regions learnt over the training sets. As discussed above, canonical correlation analysis can be conceptually seen as a process to find mutual information (or an intersection set) of any two sets.

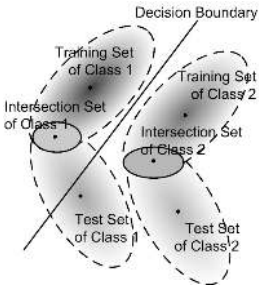


Fig. 3. Canonical Correlation Based Classification

3 SIFT Descriptor for Spatiotemporal Volume Data

Edge-based description of each plane of videos can help the method achieve more robust gesture recognition. In this section we propose a simple and effective SIFT (Scale-Invariant Feature Transform) [7] representation for a spatiotemporal data by a fixed grid. As explained, the spatiotemporal volume is broken down into three sets of orthogonal planes (XY-, YT- and XT-planes) in the method. Along each data domain, there is a finite number of planes which can be regarded as images. Each of these images is further partitioned into $M \times N$ patches in a predefined fixed grid and the SIFT descriptor is obtained from each patch (see Figure 4a). For each image, the feature descriptor is obtained by concatenating the SIFT descriptors of several patches in a predefined order. The SIFT representation of the three sets of planes is directly integrated into the proposed method in Section 2 by replacing the sets of image vectors with the sets of the SIFT descriptors prior to canonical correlation analysis. The experimental results tell that the edge-based representation generally improves the intensity-based

representation in both of the joint space-time domain (YT-, XT-planes) and the spatial domain (XY-planes).

SIFT obtained from 3D blocks. This section presents a general 3D extension of SIFT features. Traditional classifiers such as Support Vector Machine (SVM)/ Relevance Vector Machine (RVM) are applied to the video data represented by the 3D SIFT so that they can be compared with the proposed method (with SIFT) in the same data domain. Given a spatiotemporal volume representing a gesture sequence, the volume is firstly partitioned into $M \times N \times T$ tiny blocks. Within each tiny block, further analysis is done along XY-planes and YT-planes (see Figure 4b). For analysis on a certain plane, say XY-planes, derivatives along X- and Y- dimensions are obtained and accumulated to form several regional orientation histograms (under a 3D Gaussian weighting scheme). For each tiny block, the resultant orientation histograms of both planes are then concatenated to form the final SIFT descriptor of dimension 256. The descriptor for the whole spatiotemporal volume can then be formed by concatenating the SIFT descriptors of all tiny blocks in a predefined order. The spatiotemporal volume is eventually represented as a single long concatenated vector.

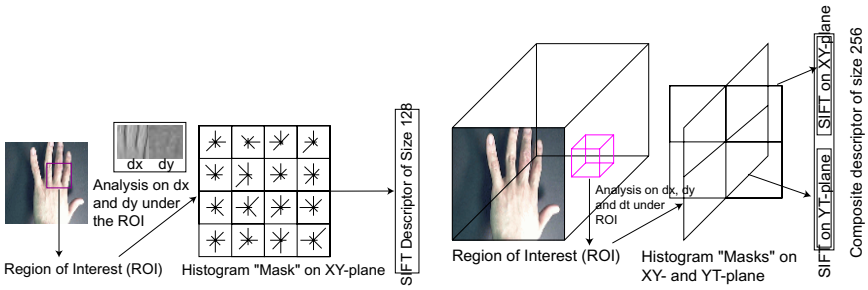


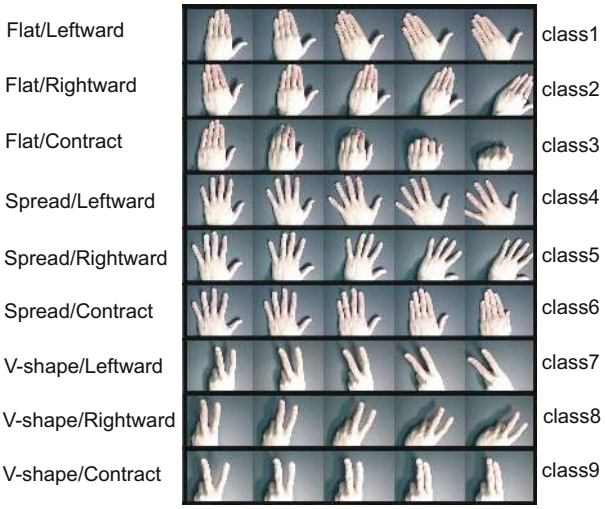
Fig. 4. SIFT Representation: (a) SIFT used in [7]. (b) SIFT from 3D blocks (refer to text).

4 Empirical Evaluation

4.1 Cambridge Hand Gesture Data Set and Experimental Protocol

We have acquired the hand-gesture data base [2] consisting of 900 image sequences of 9 gesture classes. Each class has 100 image sequences (5 different illuminations $\times$ 10 arbitrary motions of 2 subjects). Each sequence was recorded in a frame rate of 30fps and a resolution of 320×240. The 9 classes are defined by the 3 primitive hand shapes and 3 primitive motions (See Figure 5). See Figure 5c for the example images captured under the 5 different illumination settings. The

[2] The database is available upon request. Contact e-mail: tkk22@cam.ac.uk



(a) 9 gesture classes formed by 3 shapes and 3 motions.

(b) 5 illumination settings.

Fig. 5. Hand-Gesture Database

data set has temporally isolated gesture sequences which exhibit variations in initial positions, postures of hands and speed of movements in different sequences. All training was performed on the data acquired in a single illumination setting while testing was done on the data acquired in the remaining settings. The 20 sequences in the training set were randomly partitioned into the 10 sequences for training and the other 10 for the validation.

4.2 Results and Discussions

We compared the accuracy of 9 different methods:

- Applying Support Vector Machine (SVM) or Relevance Vector Machine (RVM) on Motion Gradient Orientation Images [11] (MGO SVM or MGO RVM),
- Applying RVM on the 3D SIFT vectors described in Section 3 (3DSIFT RVM),
- Using the canonical correlations (CC) (i.e. the method using $\mathcal{G}(\mathbf{P}_m^1, \mathbf{P}_q^1)$ in (2), spatiotemporal canonical correlations (ST-CC), discriminative ST-CC (ST-DCC),
- Using the canonical correlations of the SIFT descriptors (SIFT CC), spatiotemporal canonical correlations of the SIFT vectors (SIFT ST-CC), and SIFT ST-CC with the discriminative transformations (SIFT ST-DCC).

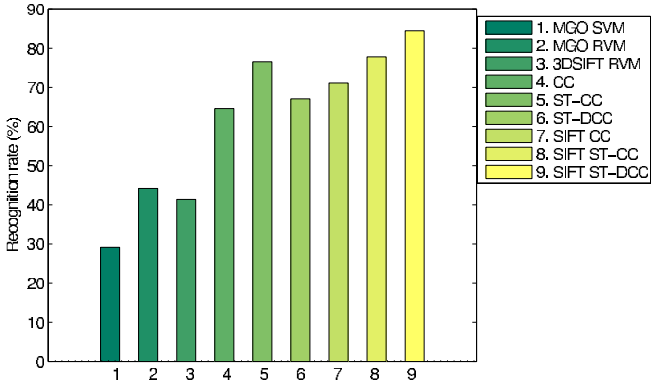


Fig. 6. Recognition Accuracy: The identification rates (in percent) of all comparative methods are shown for the plain lighting set used for training and all the others for testing

In the proposed method, the weights w^k were set up proportionally to the accuracy of the three subspaces for the validation set and Nearest Neighbor classification (NN) was done with the defined similarity functions.

Figure 6 shows the recognition rates of the 9 methods, when the plain lighting set (the leftmost in Figure 5c) was exploited for training and all the others for testing. The approaches of using SVM/RVM on the motion gradient orientation images are the worst. As observed in [11], using RVM improved the accuracy of SVM by about 10% for MGO images. However, we got much poorer accuracy than those in the previous study [11] mainly due to the following reasons: The gesture classes in this study were defined by hand shapes as well as motions. Both methods often failed to discriminate the gestures which exhibit the same motion of the different shapes, as the methods are mainly based on motion information of gestures. A much smaller number of sequences (of a single lighting condition) used in training is another reason to get the performance degradation. The accuracy of the RVM on the 3D-SIFT vectors was also poor. The high dimension of the 3D-SIFT vectors and small sample size might prevent the classifier from learning properly, as discussed. We measured the accuracy of the RVM classifier for the different numbers of the blocks in the 3D-SIFT representations (2-2-1,3-3-1,4-4-1,4-4-2 for X-Y-T) and obtained the best accuracy for the 2-2-1 case, which yields the lowest dimension of the 3D-SIFT vectors. Canonical correlation-based methods significantly outperformed the previous approaches. The proposed spatiotemporal canonical correlation method (ST-CC) improved the simple canonical correlation method by about 15%. The proposed discriminative method (ST-DCC) unexpectedly decreased the accuracy of ST-CC, possibly due to overfitting of discriminative methods. The train set did not reflect the lighting conditions in the test set. However, note that the discriminative method improved the accuracy when it was applied to the SIFT representations rather than using intensity images (See SIFT ST-CC and SIFT

Table 1. Evaluation of the individual subspace

	CC				SIFT CC			
(%)	XY	XT	YT	ST	XY	XT	YT	ST
mean	64.5	40.2	56.2	78.9	70.3	61.8	58.3	80.4
std	1.3	5.9	5.3	2.4	2.1	3.3	4.0	3.2

Table 2. Evaluation for different numbers of blocks in the SIFT representation: E.g. 2-2-1 indicates the SIFT representation where X,Y,and T axes are divided into 2,2,1 segments respectively

	2-2-1		3-3-1		4-4-1		4-4-2	
(%)	ST-CC	ST-DCC	ST-CC	ST-DCC	ST-CC	ST-DCC	ST-CC	ST-DCC
mean	80.3	80.0	78.9	83.8	80.4	85.1	75.9	83.4
std	1.9	2.5	3.6	2.7	3.2	2.8	2.4	0.7

ST-DCC in Figure 6). The proposed three methods using the SIFT representations are better than the respective three methods of the intensity images. The best accuracy was achieved by the SIFT ST-DCC at 85%.

Table 1 and Table 2 show more results about the proposed method, where all 5 experimental results (corresponding to each illumination set used for training) are averaged. As shown in Table 1 canonical correlations of the XY subspace obtained better accuracy with smaller standard deviations than the other two subspaces, but all three are relatively good compared with the traditional methods, MGO SVM/RVM and 3DSIFT RVM. Using the SIFT representation considerably improved the accuracy of the intensity images for each subspace, whereas the improvement for the joint representation was relatively small. Table 2 shows the accuracy of ST-CC and ST-DCC for the different numbers of the blocks of the SIFT representation. The best accuracy was obtained for the case of 4-4-1 for XYT (each number indicates the number of divisions along one axis). Generally, using the discriminative transformation improved the accuracy of ST-CC for the SIFT representation. Note that accuracy of the method is not sensitive about settings in number of the blocks, which is practically important.

Also, the proposed approach based on canonical correlations is computationally cheap taking computations $O(3 \times d^3)$, where d is the dimension of each subspace (which was 10), and thus facilitates efficient gesture recognition in a large data set.

5 Conclusion

A new method based on subspace has been proposed for gesture recognition under small sample size. Unlike typical classification approaches directly operating with input space, the proposed method reduces input dimension using the three sets of orthogonal planes. The method provides robust spatiotemporal volume

matching by analyzing mutual information (or canonical correlations) between any two gesture sequences. Experiments for the 900 gesture sequences showed that the proposed method significantly outperformed the traditional classifiers and yielded the best classification result using the discriminative transformations and SIFT descriptors jointly. The method is also practically attractive as it does not involve significant tuning parameters and is computationally efficient.

References

1. Björck, Å., Golub, G.H.: Numerical methods for computing angles between linear subspaces. Mathematics of Computation, 27(123) 579–594, 1973.
2. Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language. In: ECCV, pp. 390–401 (2004)
3. Darrell, T., Pentland, A.: Space-time gestures. In: Proc. of CVPR, pp. 335–340 (1993)
4. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proc. of ICCV, pp. 726–733 (2003)
5. Freeman, W., Roth, M.: Orientation histogram for hand gesture recognition. In: Int'l Conf. on Automatic Face and Gesture Recognition (1995)
6. Just, A., Rodriguez, Y., Marcel, S.: Hand posture classification and recognition using the modified census transform. In: Int'l Conf. on Automatic Face and Gesture Recognition (2006)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
8. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: Proc. of CVPR 2005, pp. 405–412 (2005)
9. Starner, T., Pentland, A., Weaver, J.: Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. 20(12), 1371–1375 (1998)
10. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. on PAMI 29(6), 1005–1018 (2007)
11. Wong, S., Cipolla, R.: Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images. In: Proc. of BMVC 2005, pp. 379–388 (2005)
12. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC (2006)
13. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR 2004, pp. 32–36 (2004)
14. Kim, T., Wong, S., Cipolla, R.: Tensor Canonical Correlation Analysis for Action Classification. In: CVPR (2007)
15. Hardoon, D., Szedmak, S., Taylor, J.S.: Canonical correlation analysis; An overview with application to learning methods. Neural Computation 16(12), 639–2664 (2004)