Discriminative human action recognition using pairwise CSP classifiers *

Ronald Poppe and Mannes Poel
University of Twente, Dept. of Computer Science, Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede, the Netherlands
{poppe,mpoel}@ewi.utwente.nl

Abstract

We present a discriminative approach to human action recognition. At the heart of our approach is the use of common spatial patterns (CSP), a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. Such a transformation focusses on differences between classes, rather than on modelling each class individually. As a results, to distinguish between two classes, we can use simple distance metrics in the low-dimensional transformed space. The most likely class is found by pairwise evaluation of all discriminant functions. Our image representations are silhouette boundary gradients, spatially binned into cells. We achieve scores of approximately 96% on a standard action dataset, and show that reasonable results can be obtained when training on only a single subject. Future work is aimed at combining our approach with automatic human detection.

1. Introduction

Automatic recognition of human actions from video is an important step towards the goal of automatic interpretation of human behavior. This understanding has many potential applications, including improved human-computer interaction, video surveillance and automatic annotation and retrieval of stored video footage.

For fully automatic recognition of human actions, we need to localize the humans in the image first. In this paper, we assume that we have access to these locations. While this might seem unrealistic, related work by Thurau [15] and Zhu *et al*. [18] shows that this detection can be performed reliably, and within reasonable time.

In the development of an action recognition algorithm, one issue is the type of image representation that is used. At one extreme, bag-of-word approaches [2] have been used. At the other extreme, pose information is used (e.g. [1]). We use a grid-based silhouette descriptor, where each cell is a histogram of oriented boundary points. This representation resembles the concept of histograms of oriented gradients (HOG, [6]), since it models the spatial relations, yet is able to generalize over local variations.

For classification, we learn simple functions that can discriminate between two classes. Our main contribution is the application of common spatial patterns (CSP), a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. After applying CSP, the first components of the transformed feature space contain high temporal variance for one class, and low variance for the other. This effect is opposite for the last components. For an unseen sequence, we calculate the histogram over time, using only a fraction (the first and last components) of the transformed space. Each action is represented by the mean of the histograms of all corresponding training sequences, which is a very compact but somewhat naive representation. A simple classifier distinguishes between the two classes. All discriminant functions are evaluated pairwise to find the most likely action class. This introduces a significant amount of noise over class labels, but works well for the given task.

We obtain competitive results on the standard action dataset introduced in [3]. The advantage of our method is that we require relatively few training samples. In fact, despite considerable variation in action performance between persons, we obtain reasonable results when training on data of a single subject. Also, we avoid retraining all functions when adding a new class, since the discriminative functions are learned pairwise, instead of jointly over all classes.

We discuss related work on action recognition from monocular video in the next section. Common spatial patterns, and the construction of the CSP classifiers, are discussed in Section 3. We evaluate our approach on a standard dataset. We summarize our experiments, and compare our results with previous work in Section 4. We conclude in Section 5.

*This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access, publication AMIDA-140), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024. We wish to thank the authors of [3] for making their dataset available.

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

2. Related work on action recognition

There has been a considerate amount of research into the recognition and understanding of human actions and activities from video and motion capture data. We discuss related work on action recognition, obtained from segmented monocular video sequences. A more comprehensive overview appears in [8].

Action recognition can be thought of as the process of classifying arbitrary feature streams, obtained from video sequences. This reveals the two main components of the problem: the feature representation and the classification. There have been many variations of either.

The choice of feature representation is important since it partly captures the variation in human pose, body dimension, appearance, clothing, and environmental factors such as lighting conditions. An ideal representation would be able to discriminate between poses, while at the same time be able to generalize over other factors. Since it is difficult to robustly obtain rich descriptors from video, often a compromise is sought in the complexity of the representation. Many approaches use retinotopic representations, where the person is localized in the image. The image observations, such as silhouette or edge representations, are conveniently encoded into a feature vector. At the other end, there is the bag-of-words approach, where the spatial dimension is ignored altogether. Feature representations that are somewhere in between these concepts, such as grid-based descriptors (e.g. [9, 16]), are currently popular. These encode the image observation locally as a bag-of-words, but preserve the spatial arrangement of these local descriptors.

When it comes to classifiers, we can generally distinguish two large classes of action recognition approaches. *Spatio-temporal templates* match unseen sequences to known action templates. These templates can take many forms. A key frame or the mean of a sequence of silhouettes over time could be templates. Slightly more advanced is the concept of Motion History Image, introduced by Bobick and Davis [4]. Here, the differences between subsequent silhouettes is used, and stored in a two-dimensional histogram. Recent work by Blank *et al*. [3] concatenates silhouettes over time to form a space-time shape. Special shape properties are extracted from the Poisson solution, and used for shape representation and classification.

The time dimension plays an important role in the recognition of actions, since there is often variation in the timing and speed in which an action is performed. Spatio-temporal templates can be considered as prototypes for a given action. Especially when using histograms, the temporal aspect is often poorly modelled.

State-based representations, the second class of action classification approach, model the temporal aspect more accurately. These methods are often represented as a graphical model, where inference is used to perform the classification. The temporal relations between different states are encoded as transition probabilities. Hidden Markov Models (HMM) have been used initially [5]. HMMs are also used by Weinland *et al*. [17] for action recognition from arbitrary, monocular views. A very similar approach using Action Nets has been taken by Lv and Nevatia [10].

Generative models try to maximize the likelihood of observing any example of a given class. For different actions that show many similarities yet have a significant intra-class variance in performance (e.g. walking and jogging), generative models do a poor job in the classification task. Another drawback of generative models is the assumption that observations are independent. Discriminative models such as conditional random fields (CRF) condition on the observation, which makes the independence assumption unnecessary. Such models can model long-range dependencies between observations, as well as overlapping features.

Recently, several discriminative alternatives have been proposed, based on CRFs. Sminchisescu *et al*. [14] use CRFs and Maximum Entropy Markov Models (MEMM) to jointly learn models for different actions from image observations or motion capture data. Wang and Suter employ factorial conditional random fields (FCRF), Quattoni *et al*. [13] use hidden conditional random fields (HCRF) that model the substructure of an action in hidden states. State-based approaches usually have a large number of parameters that need to be determined during training. This requires a sufficient amount of training data, which is not always available.

In our approach, we learn functions that discriminate between two classes. Yet, we avoid having to estimate a large number of parameters by representing actions as prototypes. These prototypes lie in a space that is transformed by applying common spatial patterns on the feature data, which are HOG-like representations of silhouette boundaries. We reduce the dimensionality of the feature representation, and select the components that maximize the variance between the two classes. For an unseen action, we pairwise evaluate all discriminant functions, where each function softly votes into the two classes. Our estimated class label corresponds to the action that received most of the votes. Even though such an approach inherently generates much noise in the classification, we show that we can accurately recognize actions, even when few training sequences are used.

3. Common spatial patterns

Common Spatial Patterns (CSP) is a spatial filter technique often used in classifying brain signals [11]. It transforms temporal feature data by using differences in variance between two classes. After applying the CSP, the first components of the transformed data have high temporal variance for one class, and low temporal variance for the other. For the last components, this effect is opposite. When transforming the feature data of an unseen sequence, the tempo-

ral variance in the first and last components can be used to discriminate between the two classes.

Consider the case that we have training sequences for two actions, a and b. Each training sequence can be seen as $n \times m_p$ matrix, where n is the number of features and m_p is number of time samples. We assume that the data is normalized in such a way that the mean of each feature is 0. Let C_a be the concatenation of the examples of action a, C_a is an $n \times m_a$ matrix. We do the same for action b to construct the matrix C_b. Now consider the matrix:

$$C = C_a C_a^T + C_b C_b^T \tag{1}$$

C is the variance of the union of the two data sets. Since C is symmetric, there exists a orthogonal linear transformation U such that $\Lambda = UCU^T$, a positive diagonal matrix. The next step is to apply the whitening transformation $\Psi = \sqrt{\Lambda}^{-1}$, which gives us $(\Psi U)C(\Psi U)^T = I$, and thus:

$$\begin{aligned} S_a &= (\Psi U) C_a C_a^T (\Psi U)^T && (2)\\ S_b &= (\Psi U) C_b C_b^T (\Psi U)^T && (3)\\ S_a + S_b &= I && (4) \end{aligned}$$

Since S_a is symmetric, there exists a orthogonal transformation D such that DS_aD^T is a diagonal matrix with decreasing eigenvalues on the diagonal. Hence DS_bD^T is also a diagonal matrix but with increasing eigenvalues on the diagonal. The CSP is the spatial transform $W = D\Psi U$ which transforms a data sequence into a sequence of dimension $2k$ such that a vector belonging to one action has high values in the first k components. For a vector of the other action, the situation is opposite. Hence, the temporal variance in these first and last components can be used to discriminate between action a and b.

3.1. CSP classifiers

Based on the CSP technique, we design discriminating functions $g_{a,b}$ for every action a and b with $a \neq b$. First we calculate the CSP transformation $W_{a,b}$ as described above. Then we apply $W_{a,b}$ to each action sequence of class a and b. Afterwards, the mean is taken over the entire sequence. This results in a single n-dimensional vector which can be considered a histogram, normalized for the length of the sequence. Next, we calculate the mean of these training vectors for action a and b, $\bar{a}$ and $\bar{b}$, respectively. In order to compute $g_{a,b}(x)$ for a new action sequence x, we use the same procedure and first apply $W_{a,b}$ to x. We then calculate the mean over time over all components, which gives a vector x' of length n. Finally, $g_{a,b}(x)$ is defined as follows:

$$g_{a,b}(x) = \frac{\|\bar{b} - x'\| - \|\bar{a} - x'\|}{\|\bar{b} - x'\| + \|\bar{a} - x'\|} \tag{5}$$

Evaluation of a discriminant function gives an output in the $[-1, 1]$ interval. Note that $g_{a,b} + g_{b,a} = 0$. Now the action sequence is classified by evaluating all discriminant functions between pairs of a and b over all actions:

$$g_a(x) = \sum_{a \neq b} g_{a,b}(x) \tag{6}$$

Since each action class appears in the exact same number of discriminative functions, the classification of x is the action a for which $g_a(x)$ is maximal. We also evaluate the discriminant functions in which the actual class does not appear. This introduces a large component of noise into the voting. However, actions that show more similarities with the unseen sequence will receive more mass in the voting.

4. Experimental results

We evaluate our approach on a publicly available dataset, which is briefly described in Section 4.1. Our image representation is discussed in Section 4.2. We present the setup of our experiments, and our obtained results in Section 4.3. A discussion of the results, and comparison with related work are given in Sections 4.4 and 4.5, respectively.

Figure 1. Example frames from the Weizmann dataset. Different subjects perform actions bend, jack, jump, p-jump, run, side, skip, walk and wave1. Wave2 is not shown due to space limitations.

4.1. Human action dataset

For the evaluation of our approach, we used the Weizmann action dataset [3]. This set consists of 10 different actions, each performed by 9 different persons (see also Figure 1). For person *Lena*, additional trials appear for the run, skip and walk action. We decided to leave these out, to have a balanced set. This also allows for direct comparison of our results to those previously reported on the dataset. Note that

our approach also works for unbalanced sets. The skip action was originally not present in the set, and we present results both with and without the skip action.

Each trial takes approximately 2.5 seconds. There is considerable intra-class variation due to different performances of the same action by different persons. Most notably, the run, skip and walk actions are performed either from left to right, or in opposite direction. The trials are recorded from a single camera view, against a static background, with minimal lighting differences. Binary silhouette masks are provided with the dataset.

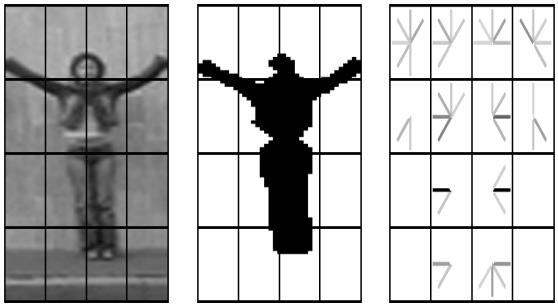
4.2. HOG-like silhouette descriptors

Similar to recent work on action recognition [9, 15, 16], we use a grid-based approach, see also Figure 2. Given an extracted silhouette, we determine the minimum enclosing bounding box, which determines the region of interest (ROI). We add space to make sure the height is 2.5 times the width. Next, we divide the ROI into a grid of 4×4 cells. Within each cell, we calculate the distribution of silhouette gradients, which we divide over 8 bins that each cover a 45° range. This idea is similar to that of histogram of oriented gradients (HOG, [6]) but our implementation is a simplification at a number of levels. First, we do not apply a Gaussian filter to enhance the edges. Second, we do not use overlapping cells, which significantly reduces the size of our descriptor. Third, and most important, we only take into account the silhouette outline, thus discarding the internal edges. The final 128-dimensional descriptor is a concatenation of the histograms of all cells, normalized to unit length to accommodate variations in scale.

Due to this normalization, and the relatively high dimensionality compared to the number of data points in a trial, the covariance over a trial might be nearly singular in some cases. We avoid this by applying PCA, and selecting the 50 first components. These explain approximately 75% of the variance, depending on the person that is left out. See the next section for details regarding this process.

4.3. Results

We evaluate our method using leave-one-out cross-validation (LOOCV), where each of the 9 folds corresponds to all trials of the corresponding person. Specifically, this gives us 80 training sequences per fold, 8 for each of the 10 actions. First, we calculate the PCA transformation over all training sequences, and project the silhouette descriptors down on the first 50 components. Next, we learn all discriminant functions $g_{a,b}$, between all pairs of actions a and b ($1 \leq a, b \leq 10, a \neq b$). Specifically, we use the first and last $k = 5$ components in the transformation, which gives us action prototypes vectors of dimension 10. We experimented with other values for k but found no improvement for $k > 5$. For each of the trials of the person whose sequences were left out, we evaluate all these discriminant functions. Each of these evaluations softly votes over class a and b. In our final classification, we select the class that received the highest voting mass.



(a) (b) (c)

Figure 2. Silhouette descriptor, (a) image, (b) mask and (c) the boundary orientations, spatially binned into cells. Normal vectors are shown for clarity.

We performed the LOOCV, and obtained a performance of 95.56%. In total, 4 trials were misclassified. The skip action of subject *Daria* was classified as jumping, the skip action of subject *Ido* was classified as running. Also, the jump action of subject *Eli* and the run action of subject *Shahar* were both classified as walking.

To be able to compare our results with those reported in previous studies, we also left out the skip class. This resulted in a performance of 96.30%. Again, the jump action of subject *Eli* and the run action of subject *Shahar* were classified as walking. In addition, the wave1 action of Subject *Lyova* was misclassified as wave2.

4.4. Discussion

The baseline for the full dataset is 10%, and 11.11% when the skip action is left out. Obviously, our results are well above these baselines, and show that we can achieve good recognition, even when single action prototypes of dimension 10 are used. Also, it shows that intra-class variations can be handled, without modelling the variance between different subjects. This gives the impression that we can even train our classifiers with less training data. Note here that training for all actions and all training subjects takes well under 1 second. To verify this hypothesis, we evaluated the performance of our approach using different numbers of subjects in the training set. For each number m, we present the results as averages over all $8!/(m!(8-m)!)$ combinations of training subjects. Table 1 summarizes these results, both using all actions, and with the skip action omitted.

Clearly, performance decreases with a decreasing amount of training data. But even when only a few subjects

Table 1. Classification performance of our CSP classifier on the Weizmann dataset, using different numbers of training subjects. Combinations is the evaluated number of subsets of subjects.

Subjects	Combinations	All actions	Skip omitted
1	8	64.72%	69.14%
2	28	77.82%	83.82%
3	56	81.83%	88.98%
4	70	84.60%	90.85%
5	56	86.63%	92.44%
6	28	89.01%	93.87%
7	8	91.39%	94.91%
8	**1**	**95.56%**	**96.30%**

are used for training, the results are reasonable. Especially for the evaluations with very few training subjects, we expect that the variation in the direction of movement of the run, skip and walk sequence will have a significant impact on the results. Even though we do not model the movement in the image, changing the direction of movement results in image observations to be mirrored. Of course, this results in very different silhouette descriptors. Still, our approach can cope with these variations to some extent.

Both the feature representation and the classifier have an important impact on the performance. To measure the added value of using CSP, we performed an additional experiment where we did not transform the feature space. Instead, we directly took the first 10 components of the PCA. For each training sequence, we calculated the histogram by taking the mean of the feature vector over time, which resulted in a 10-dimensional vector. We determined the prototype for each action by averaging over these histograms. Again, we used Equation 5 and 6 to determine the class estimate. We achieved a performance of 77.78% for all actions, and 85.19% with the skip action omitted. When we used the first 50 PCA components, the performance slightly increased to 80.00% for all actions, while the performance without skip remained the same. A closer look at the misclassifications shows confusion between run, skip and walk, along with some incidental confusions. It thus becomes clear that the use of CSP is advantageous over a feature representation without CSP transform.

4.5. Comparison with related research

There have been several reports of results on the same data set. We review these results, and point out differences with our work. Such comparisons reveal the advantages and disadvantages of one method over the other.

Niebles and Fei-Fei [12] achieve a 72.80% score over 9 actions. Spatial and spatio-temporal interest points are sampled, and combined into a constellation. Action classification is performed by taking a majority vote over all individually classified frames. No background segmentation or localization is needed. This makes their approach more robust than ours. Recent work by Thurau [15] uses HOG-descriptors for both detection and action classification. No background segmentation is used, but centered and aligned training data is needed. For classification, n-grams of action snippets are used. With all 10 actions, and 90 bi-grams, a performance of 86.66% was achieved.

The work of İkizler and Duygulu [9] does not require background segmentation, but localization is assumed. A large number of rotated rectangular patches is extracted, and divided over a 3×3 grid, forming a histogram of oriented rectangles. A number of settings and classification methods was evaluated on the dataset without the skip action. All actions were classified correctly when using Dynamic Time Warping. This requires the temporal alignment of each unseen sequence to all sequences in the training set, which is computationally expensive. Using one histogram per sequence, 96.30% was scored. Again, this requires comparison to all training sequences. For comparison, we calculated the performance of our descriptor using global histograms and 1-nearest neighbor using Euclidian distance, and with the skip action left out. This resulted in 96.30% performance, a similar score.

Other works require background subtraction, and use the masks that are provided with the dataset. Wang and Suter [16] score 97.78% over all 10 actions. Raw silhouette values are used, and long-term dependencies between observations are modelled in their FCRF. When small blocks of pixels are regarded, thus effectively reducing the resolution, performance decreased. For 4×4 blocks and 8×8 blocks, scores were obtained of 92.22% and 77.78%, with descriptor sizes 192 and 48, respectively. Kernel PCA was used to reduce the dimensionality, but the dimension of the projected space was not reported. In contrast, we start with a 128-dimensional silhouette descriptor, and perform the classification using only 10 components. Moreover, our training requirements are much lower. On the other hand, FCRFs are able to model complex temporal dynamics.

Blank *et al*. [3], and more recently [7], create space-time volumes for each sequence by concatenating silhouettes over time. Special shape properties are extracted from the Poisson solution, and used for shape representation and classification. Classification is performed using 1-nearest neighbor. Performance was measured on smaller subsequences, which makes direct comparison impossible. Their performance is 99.61% without skip, and 97.83% on the full dataset. Note that the use of subsequences partly solves the issue of partial silhouettes when a person enters the image, which is that case of half of the actions. Batra *et al*. [2] use a bag-of-words approach, with small space-time silhouette patches as feature points. Classification on the dataset without skip results in 82.72% performance using 1-nearest neighbor, and 91.36% using a logistic regressor.

5. Conclusion & future work

In this paper, we performed action recognition on a standard dataset. Oriented silhouette boundaries are encoded as a histogram of orientated gradients within in cells of a grid. The main contribution of our work is the application of common spatial patterns (CSP). This is a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. Such a transformation focusses on differences between classes, rather than on modelling each class individually. For unseen sequences, we calculate the histogram over time, using only the first and last 5 components of the transformed space. Each action is represented by the mean histograms of all corresponding training sequences. Simple discriminant functions are evaluated pairwise to find the most likely action class.

Using the complete Weizmann action dataset, we obtained a recognition performance of 95.55%. With the skip action left out, as in previous research, our performance is 96.30%. These results are competitive, and we showed that we can even obtain reasonable results with only a few training subjects. Moreover, training complexity is low.

Future work will be aimed at performing automatic human detection, such as in [18]. Also, we plan to investigate action prototypes that can model temporal characteristics. In addition, we would like to investigate other classification schemes, to avoid the noise that is introduced when using pairwise evaluation. A final point to address is the temporal segmentation of the action.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[2] D. Batra, T. Chen, and R. Sukthankar. Space-time shapelets for action recognition. In *Proceedings of the Workshop on Motion and Video Computing (WMVC'08)*, Copper Mountain, CO, January 2008.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 2*, pages 1395–1402, Beijing, China, October 2005.

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

[5] M. Brand, N. Oliver, and A. P. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 994–999, San Juan, Puerto Rico, June 1997.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1*, pages 886–893, San Diego, CA, June 2005.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(12):2247–2253, January 2007.

[8] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions On Systems, Man, And Cybernetics (SMC) - Part C: Applications And Reviews*, 34(3):334–352, August 2004.

[9] N. İkizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *Human Motion: Understanding, Modeling, Capture and Animation*, number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil, October 2007.

[10] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[11] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, 110(5):787–798, May 1999.

[12] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[13] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(10):1848–1852, October 2007.

[14] C. Sminchisescu, A. Kanaujia, and D. N. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):210–220, November 2006.

[15] C. Thurau. Behavior histograms for action recognition and human detection. In *Human Motion: Understanding, Modeling, Capture and Animation*, number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil, October 2007.

[16] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[17] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[18] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1491–1498, New York, NY, June 2006.