

Action spotting exploiting the frequency domain

Irene Kotsia

School of Electronic Engineering and Computer Science
Queen Mary University of London, Mile end Campus, UK, E14NS

irene.kotsia@eeecs.qmul.ac.uk

Vasileios Argyriou

Faculty of Computing, Information Systems and Mathematics
Kingston University London, Kingston upon Thames, Surrey KT1 2EE

Vasileios.Argyriou@kingston.ac.uk

Abstract

In this paper we present a novel method for action spotting in a video sequence, that employs the 3D Fourier transform in order to define a 3D gradient correlation function operating at the frequency domain. In that way we achieve invariance to spatiotemporal variations and also to frame reordering. One of the most attractive features of the proposed scheme is its high degree of computational efficiency and the fact that it can be implemented by fast transformation algorithms in the frequency domain. Experiments conducted to a publicly available database demonstrated the superiority of the proposed method in terms of more accurate estimations and significantly lower computational complexity.

1. INTRODUCTION

Action spotting refers to the detection and spatiotemporal localization of an action executed by a person over a small period of time, in a video sequence. Its applications span several areas, from video query in a reference video database to surveillance and tracking applications.

Efficient action spotting faces a lot of problems mainly caused by appearance issues. More precisely, changes in clothing or environmental lighting conditions lead to different image intensities, producing in that way varying patterns. Moreover, the presence of clutter adds extra difficulty. For example background clutter (appearing when the background is either dynamic or very complicated) or foreground clutter (occlusions) significantly affect the appearance of the extracted pattern, thus making the action spotting procedure more complex.

A limited number of works have been proposed in the past years to tackle the above mentioned problems. In

[20] the authors exploit the curvature scale space of silhouettes, offering in that way invariance to noise and significant shape corruption. In a more recent approach [6] the authors propose a compact local descriptor based on visual space-time oriented energy measurements. The developed descriptor allows for the comparison of the underlying dynamics of two space-time video segments irrespective of spatial appearance, such as differences induced by clothing, and with robustness to clutter. The authors employ the Fourier transform in order to achieve fast calculation of the convolution of the input imagery with a set of Gaussian third derivative filters. The result is used to calculate an energy function, used for template matching in order to achieve action spotting.

The differences of our work with the approach presented in [6] are threefold. First, we use the Fourier transform to define 3D gradient correlation. Second, we provide space and time scale invariance using the proposed 3D gradient correlation, while in [6] the authors provide invariance through the proposed marginalized energy function. Third, we also deal with action patterns whose frames need reordering, something that is not the case for the method presented in [6]. Furthermore, in terms of performance the proposed method has lower computational complexity allowing faster action spotting.

During the past decade, human action recognition has attracted the scientific community's interest due to its wide applications in human computer interfaces, video surveillance etc. A great amount of research has been conducted towards human action recognition of a small set of basic activities under controlled conditions and can be distinguished in three categories: those that use articulated models, those that use spatio-temporal features and the last category that includes the methods that use motion templates.

In more detail, the first category includes methods that

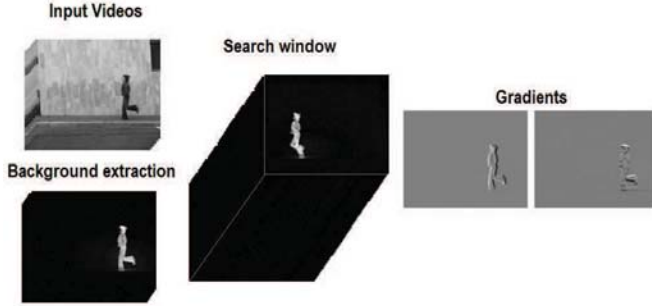


Figure 1. Overview of the proposed method

estimate the parameters of articulated models of human bodies and apply vector-based methodologies on the acquired parameters. In [17] the authors build a model of appearance of each person in every video and track it through time. The 3D motion sequence is then synthesized and matched against a collection of annotated motion capture data using a simple Support Vector Machines (SVMs) system. An articulated model is also used in [22], where 2D and 3D shape cues such as silhouettes, motion residue and skeleton curves are used to track the model through time using multiple cameras.

The above mentioned representations are invariant to viewpoints, since they fit complete models to the human body. They are also not affected by scale changes, as the relative position of the model’s joints remains the same throughout the motion. However, the procedure of fitting an articulated model using 2D data, acquired either from one or multiple cameras, is highly dependent on the initialization parameters. It is also more likely to fail when occlusion or clutter are present in the video. Moreover, the construction of an entire model for each human body for each video frame is exceptionally computationally complex, something extremely undesirable.

The second category is based on the 2D tracking of landmark points, such as body limbs, face etc on the image plane. Salient points that are distinctive for an action, calculated by measuring energy, are detected and used either in a bag-of-words approach [15] or to form cuboids in time [18]. The curvatures of specific human joint points are also used in [16]. A variety of different detector spatio-temporal points are also combined for better features extraction in [12]-[24]. In all the above methods, the acquired trajectories are subsequently modelled and classified using classical vector-based methodologies such as Hidden Markov Models (HMMs) and SVMs. The use of point trajectories constitutes the approach view invariant. However, accurate positions of the points are difficult to estimate, especially in the presence of occlusion, clutter and loose clothing.

The third category contains methods that rely on view-based representations, extracting templates to achieve ac-

tion recognition. Early approaches utilize holistic representations such as temporal images, calculating motion history images [4], optical flow [25] or action shapes [8]. However, the need for more efficient representations has resulted in the extraction of local features estimated on spatiotemporally salient points, that can be detected in a more robust way, despite illumination and small viewpoint variations. Thus, histograms of Oriented Gradients or of optical flow were used [5][13]. Lately, the correlation between temporal templates has been also thoroughly studied [11][13].

In this paper we present a novel and efficient method that spots an action in a reference video sequence. The method employs the 3D Fourier transform in order to define a 3D gradient correlation in the frequency domain. The proposed method is invariant to space and time changes and efficiently spots action patterns even in cases where there is frame reordering. Experiments conducted with a publicly available database demonstrated the superiority of the proposed method in terms of more accurate estimations and significantly lower computational complexity.

The remainder of this paper is organized as follows. We define the correlation in the frequency domain in Section 2 and extend it for action spotting in Section 3. Results are presented in Section 4. Conclusions are drawn in Section 5.

2. MOTION ESTIMATION IN THE FREQUENCY DOMAIN

2.1. Phase Correlation

Baseline phase correlation operates on a pair of images or, more commonly a pair of co-sited rectangular blocks f_t and f_{t+1} of identical dimensions belonging to consecutive frames or fields of a moving sequence sampled at $t, t + 1$. The estimation of motion relies on the detection of the maximum of the cross-correlation function between f_t and f_{t+1} . Since all functions involved are discrete, cross correlation is circular and it can be carried out as a multiplication in frequency domain using fast implementations. The correlation surface is defined as:

$$c_{t,t+1}(k, l) = F^{-1} \left(\frac{F_t^* F_{t+1}}{|F_t * F_{t+1}|} \right) \quad (1)$$

where F_t and F_{t+1} are respectively the two-dimensional discrete Fourier transforms of f_t and f_{t+1} , F^{-1} denotes the inverse Fourier transform and $*$ denotes complex conjugate. The co-ordinates (k_m, l_m) of the maximum of the real-valued array $c_{t,t+1}$ can be used as an estimate of the horizontal and vertical components of motion at integer pixel precision between f_t and f_{t+1} as follows:

$$(k_m, l_m) = \arg \max Re \{c_{t,t+1}(k, l)\} \quad (2)$$

where $Re\{\}$ is the real part of the complex phase correlation surface array. It is also worth mentioning similar methods

operating in the frequency domain on a square lattice presented in [2, 7, 10, 21].

2.2. Gradient Correlation

At each pixel location $f(x, y)$ of each frame discrete approximations to the horizontal and vertical components of the spatial gradient are obtained using central difference, i.e. $g_t^h(x, y) = f_t(x + 1, y) - f_t(x - 1, y)$ and $g_t^v(x, y) = f_t(x, y + 1) - f_t(x, y - 1)$. The two terms are combined in a unified complex representation of the form $g_t(x, y) = g_t^h(x, y) + jg_t^v(x, y)$, which retains dense magnitude and phase information. It should be noted that this representation is independent of the approximation to the gradient operator. In this work, standard Phase Correlation is replaced by gradient cross-correlation (*GC*) defined as follows

$$GC_{t,t+1}(k, l) = F^{-1} (G_t^* G_{t+1}) \quad (3)$$

where G_t and G_{t+1} are respectively the 2D FFTs of g_t and g_{t+} . The location of the maximum corresponds to the shift between the frames and it is given by (2).

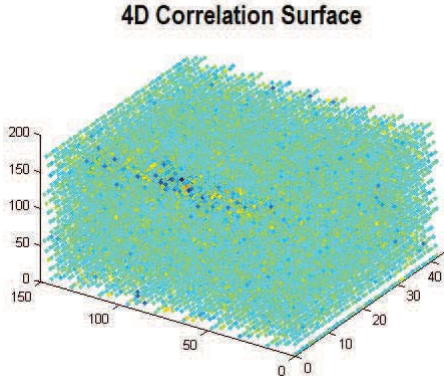


Figure 2. The obtained 4D correlation surface

2.3. Sub-pixel accuracy

Sub-pixel accuracy of motion measurements is obtained by separable-variable fitting performed in the neighbourhood of the maximum using one-dimensional quadratic functions [1, 3]. Using the notation in (2) above, the location of the maximum of the fitted function provides the required sub-pixel motion estimate (dx, dy) . For example fitting a parabolic function horizontally yields a closed-form solution for the horizontal component of the motion estimate dx as follows:

$$dx = \frac{c_{t,t+1}(k_m + 1, l_m) - c_{t,t+1}(k_m - 1, l_m)}{2(2c_{t,t+1}(k_m, l_m) - c_{t,t+1}(k_m + 1, l_m) - c_{t,t+1}(k_m - 1, l_m))}$$

The fractional part dy of the vertical component can be obtained in a similar way. Instead of a quadratic function fitting, a Gaussian or a ‘sinc’ function may also be used [7].

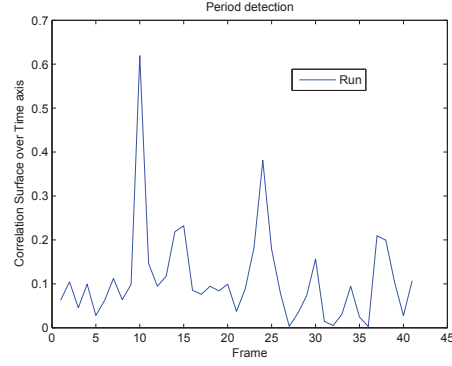


Figure 3. The obtained correlation surface over the time axis

3. Action spotting in the frequency domain

In this section Gradient Correlation (*GC*) is extended to 3D space allowing automatic action spotting and frame ordering. Let $s(x, y, t_s)$ and $q(x, y, t_q)$ be two sequences with s containing a period of a given action, while in q unknown actions are repeated. The proposed method automatically spots the given action and applies frame ordering.

In the first stage of the proposed algorithm background extraction is applied using adaptive Gaussian mixture models and a bounding box around the foreground is estimated. The same procedure is applied to both sequences and we obtain $s_b(x, y, t_s)$ and $q_b(x, y, t_q)$ corresponding to the foreground areas of action (see Fig. 1). In order to perform action spotting in the frequency domain 3D Gradient Correlation is applied on the unequal size sequences $s_b(x, y, t_s)$ and $q_b(x, y, t_q)$:

$$c_{s,q}(k, l, n) = FFT3^{-1} \left(\frac{FFT3(s_b(x, y, t_s)) * FFT3(q_b(x, y, t_s))}{|FFT3(s_b(x, y, t_s)) * FFT3(q_b(x, y, t_s))|} \right) \quad (5)$$

Nevertheless, this has the obvious disadvantage of requiring a gradient correlation operation to be performed between two sequences of unequal sizes, i.e. $t_s < t_q$. One straightforward way to go round this problem would be to increase the size of the action sample $s_b(x, y, t_s)$ by a factor in each side of the time axis by extrapolative padding, i.e., by symmetric insertion of zero frames or mid-grey values to the unknown time locations outside the sample action box until the latter assumes equal dimensions to the $q_b(x, y, t_q)$ sequence. Furthermore, it may be beneficial to carry out this operation in the frequency domain by interpolative upsampling of S_b , which is the Fourier transform of $s_b(x, y, t_s)$. Linear interpolation may be used to obtain \tilde{S}_b , whose dimensions are now identical to Q_b , which is the Fourier transform of $q_b(x, y, t_s)$, and hence allowing 3D

Gradient Correlation to be possible:

$$c_{s,q}(k, l, n) = FFT3^{-1} \left(\frac{\tilde{S}_b^* Q_b}{|\tilde{S}_b^* Q_b|} \right) \quad (6)$$

Interpolative upsampling in the frequency domain has the obvious practical advantage that the Fourier transform of the S_b sequence requires far less computations than otherwise. Furthermore, the use of background extraction provides the additional advantage of avoiding the artificial edges occurring at the transition boundary between actual data and extrapolated data. Indeed, actual data are zero in the background areas of the frames while extrapolated data are zero by definition.

The obtained correlation 4D surface (see Fig. 2) contains multiple peaks indicating the repetitions of the action during the whole sequence. Since the peaks are expected to be on the time axis the x and y axis can be removed by applying summation according to

$$\tilde{c}_{s,q}(n) = \sum_k \sum_l c_{s,q}(k, l, n) \quad (7)$$

An example of a correlation function on the time axis is shown in Fig. 3, where three peaks are observed. In this example we expected to have three peaks since only three repetitions are present.

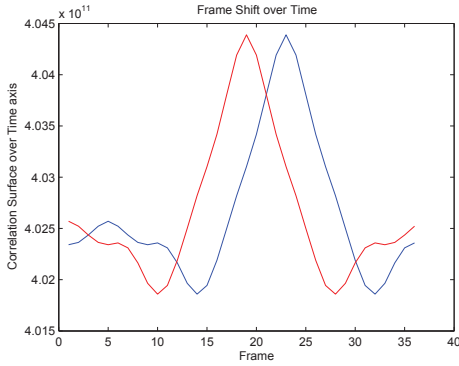


Figure 4. The obtained shift over time

In the second stage of the proposed method, frame ordering is applied in the frequency domain. Since the action spotting is positive and the a period is obtained frame ordering is a pre-required step before action recognition is performed. In this stage again 3D Gradient Correlation is applied but in this case zero-padding or interpolation in the frequency domain are not required.

$$c_{p_s,q}(k, l, n) = FFT3^{-1} \left(\frac{S_b^* Q_{p_b}}{|S_b^* Q_{p_b}|} \right) \quad (8)$$

where Q_{p_b} is one period of the spotted action. In order to obtain the frame shift summations over the x and y axis are

applied and the location of the peak indicates the required time shift to order automatically the frames. An example is shown in Fig. 4.

3.1. Space and Time scale invariance

One of the problems that it may be noticed is the sensitivity of the proposed method to object and time scaling. In order to reduce these problems the extended version of Gradient Correlation to scaling is utilised. If the video sequence $s_b(x, y, t)$ is a replica of $q_b(x, y, t)$ scaled by the factors (a, b, c) , they are related by

$$s_b(x, y, t) = q_b(ax, by, ct) \quad (9)$$

The Fourier transforms are given by

$$S_b(k, l, n) = \frac{1}{|abc|} Q_b(k/a, l/b, n/c) \quad (10)$$

and by taking logarithms:

$$S_b(\log k, \log l, \log n) = A Q_b(\log k \log a, \log l - \log b, \log n \log c) \quad (11)$$

If logarithmic axes are used, the scaling factor can be found as a shift in the frequency domain [9, 19, 23].

4. Experimental Results

For the experiments we used the well-known database for action recognition, namely the Weizmann [8] dataset. More precisely, the Weizmann Action Dataset depicts 9 subjects performing 10 different activities: “run”, “walk”, “skip”, “jumping-jack” (or shortly “jack”), “jump-forward-on-two-legs” (or “jump”), “jump-in-place-on-two-legs” (or “pjump”), “gallop sideways” (or “side”), “wave-two-hands” (or “wave2”), “wave-one-hand” (or “wave1”) and “bend”. Each video sequence is of resolution 144×160 pixels.

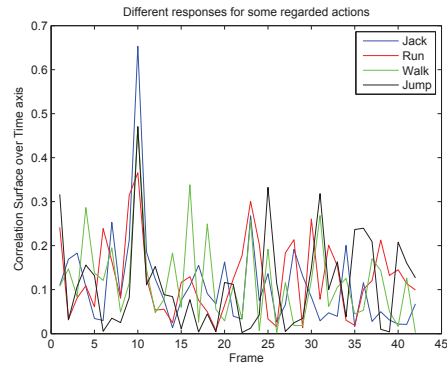


Figure 5. Different responses of correlation surfaces over the time axis for “jack” from Weizmann dataset.

Action recognition

Regarding action recognition we first calculated the value of the correlation surface of an action with a sample period of every available class. In that way we acquired a set of responses, each one corresponding to one of the available classes. The response with the greater value corresponded to the class that was assigned to the sample under examination. An example of the different responses acquired when a “jack” sample period was used from the Weizmann database, for all available actions regarded, is shown in Fig. 5. The action recognition accuracy was equal to 100% for the Weizmann dataset.

Period detection

We subsequently conducted experiments in order to find the periods of action inside a video sequence. To this end, we isolated the frames corresponding to one period of action from one video sequence and searched for them in another video sequence. In Fig. 6 we present the acquired results for the actions “bend” and “jack” from the Weizmann database. Indeed, as can be seen, several peaks appear in the graphic plot. Each peak corresponds to the middle frame of a period, thus indicating the number of periods existing in the video sequence. For the “jack” six peaks are observed. In this example we expected to have three peaks since only three repetitions are present. This can be explained by the fact that in the specific action the first half of the period is almost identical with the second half (in the inverse order). For the “bend” action only one period is present therefore only one peak is observed, while a second one just starts to appear again due to the symmetry of the action.

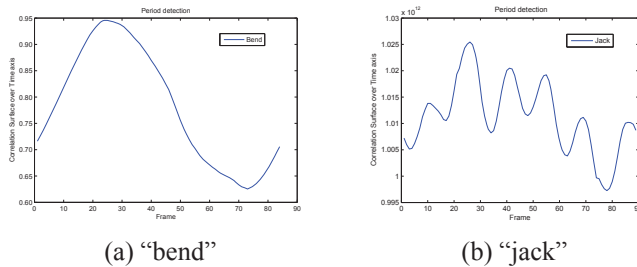


Figure 6. Period detection for a)“bend” and b)“jack” actions from Weizmann dataset.

Time shifting

Subsequently, we studied the effect of time shifting, i.e. the ability of the proposed method to correctly localize and recognize actions when their periods appear in a different time moment. We regard this case to result from a time shifting by a number of frames. As can be seen in Fig. 1 the action whose correlation surface over the time axis is depicted in the second row (“bend” and “walk” from the

Weizmann database) is the same with the response depicted in the first row, shifted by a number of frames N (in that case equal to 4 and 10). That means that the action in the second video sequence started N frames later than that from the first video sequence.

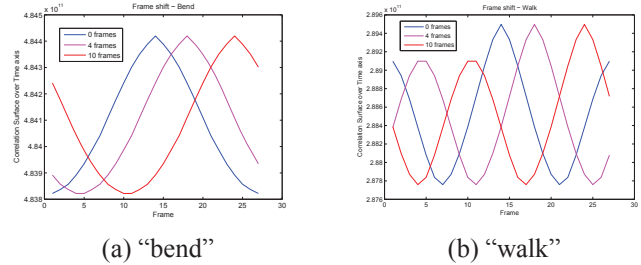


Figure 7. Effect of time shifting in bend and walk from the Weizmann database.

Symmetry in actions

The proposed method also correctly identifies symmetry parts in an action achieving in that way inversion-invariant. In Fig. 8 the action “walk” from the Weizmann database is considered. For that action, the second half of the period can be considered as the inversed version of the top half, something that is evident in the estimated correlation surface.

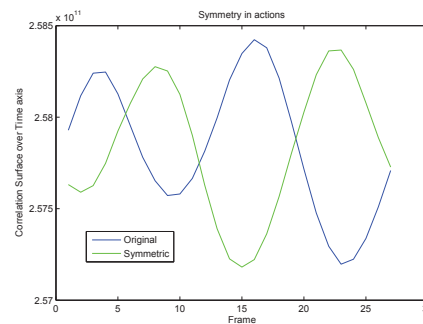


Figure 8. Spotting symmetry in video sequence.

5. CONCLUSIONS

In this paper we presented a novel and efficient method to perform action spotting in a video sequence. The method defines a 3D gradient correlation function using the 3D Fourier transform, that is invariant to spatiotemporal changes, as well as to frame reordering. Experiments conducted in a publicly available database verified the superiority of the proposed method in terms of more accurate estimations and significantly lower computational complexity.

References

- [1] I. Abdou. Practical approach to the registration of multiple frames of video images. *Proc. SPIE Conf. Vis. Comm. and Image Proc.*, 3653:371–382, Jan 1999. 3
- [2] V. Argyriou and T. Vlachos. Using gradient correlation for sub-pixel motion estimation of video sequences. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc., ICASSP04*, pages 329–331, 2004. 3
- [3] V. Argyriou and T. Vlachos. A study of sub-pixel motion estimation using phase correlation. *British Machine Vision Association, 17th BMVC*, Sept 2006. 3
- [4] A. Bobick and J. Davis. The recognition of human movements using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 2
- [5] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
- [6] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. *CVPR Recognition*, pages 1990–1997, 2010. 1
- [7] H. Foroosh, J. Zerubia, and M. Berthod. Extension of phase correlation to sub-pixel registration. *IEEE Trans. Image Processing*, 11(3):188–200, 2002. 3
- [8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2006. 2, 4
- [9] L. Hill and T. Vlachos. On the estimation of global motion using phase correlation for broadcast applications. *Proceedings IPA-1999 (IEE International Conference on Image Processing and Its Applications)*, 1:721–725, 1999. 4
- [10] W. Hoge. A subspace identification extension to the phase correlation method. *IEEE Trans. Med. Imag.*, 22(2):277–280, Feb 2003. 3
- [11] T.-K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1415–1428, August 2009. 2
- [12] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos “in the wild”. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
- [13] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
- [14] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [15] A. Oikonomopoulos and M. Pantic. Human body gesture recognition using adapted auxiliary particle filtering. *Advanced Signal and Video Based Surveillance*, pages 441–446, 2007. 2
- [16] V. Parameswaran and R. Chellappa. Human action-recognition using mutual invariants. *Computer Vision and Image Understanding*, 98(2):294–324, 2005. 2
- [17] D. Ramanan and D. A. F. and A. Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 2007. 2
- [18] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
- [19] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996. 4
- [20] M. C. Roh, B. Christmas, J. Kittler, and S. W. Lee. Gesture spotting in low-quality video with features based on curvature scale space. *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006. 1
- [21] H. Stone, M. Orchard, E. Chang, and S. Matrices. A fast direct fourierbased algorithm for subpixel registration of images. *IEEE Trans. Geo. and Rem. Sensing*, 39(10):2235–2243, Oct 2001. 3
- [22] A. Sundaresan and R. Chellappa. Multicamera tracking of articulated human motion using shape and motion cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):2114–2126, September 2009. 2
- [23] G. Thomas. Television motion measurement for datv and other applications. *BBC Res. Dept. Rep., No. 1987/11*, 1987. 4
- [24] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2
- [25] L. Wolf, H. Jhuang, and T. Hazan. Modeling appearances with low-rank svm. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007. 2