A Novel Human Action Representation via Convolution of Shape-Motion Histograms

Teck Wee Chua and Karianto Leman

Institute for Infocomm Research,
A*STAR (Agency for Science, Technology and Research), Singapore
{tewchua,karianto}@i2r.a-star.edu.sg

Abstract. Robust solutions to vision-based human action recognition require effective representations of body shapes and their dynamics. Combining multiple cues in the input space can improve the recognition task. Although conventional method such as concatenation of feature vectors is straightforward, it may not sufficiently encapsulate the characteristics of an action. Inspired by the success of convolution-based reverb application in digital signal processing, we propose a novel method to synergistically combine shape and motion histograms via convolution operation. The objective is to synthesize the output (action representation) which carries the characteristics of both source inputs (shape and motion). Analysis and experimental results on the Weizmann and KTH datasets show that the resultant feature is more efficient than other hybrid features. Compared to other recent works, the feature that we used has much lower dimension. In addition, our method avoids the need for determining weights manually during feature concatenation.

1 Introduction

There has been a surge, in recent years, towards the study of human action recognition because it is fundamental to many computer vision applications such as video surveillance, human-computer interface, and content-based video retrieval. The search results retrieved from the search engine upon the keyword 'human action recognition' is astonishing. For instant, Google search engine returned 27,800,000 results as on 9 August, 2013. While human can recognize an action in a seemingly effortless fashion, the solutions using computer have, in many cases, proved to be immensely difficult. One open problem is the choice of optimal representations for human actions. Ideally, the representation should be robust against inter/intra variations, noises, temporal variations, and sufficiently rich to differentiate huge number of possible actions. Practically, such representation does not exist.

Recent approaches can be categorized into local and global representations. Local representation encodes the image or video frames as a collection of local patches. Common local representation includes spatio-temporal interest points such as 3D Harris, cuboid, Hessian, Dense etc. Usually the interest points are extracted at different spatial and temporal scales. Laptev and Lindeberg [1] proposed to extend Harris corner detector to the third dimension. Dollár *et al.* [2]

C. Gurrin et al. (Eds.): MMM 2014, Part I, LNCS 8325, pp. 98–108, 2014.

cuboid detector is based on temporal Gabor filter. Willems *et al.* [3] measured the saliency with the determinant of the 3D Hessian matrix. For global representation, the region-of-interest (obtained by tracking or background subtraction) is encoded as a whole. In other words, the entire human figure is considered. Silhouette, optical flow, edge and space-time volume fall into this category. Efros *et al.* [4] used blurred optical flows to recognize the actions of small human figures. Blank *et al.* [5] stacked silhouettes over a sequence to form space-time volumes. Poisson equation was used to compute local space-time saliency and orientation features. Ikizler *et al.* [6] extended the motion descriptor of Efros by using spatial and directional binning and then combined it with line shape descriptor. Following that, they proposed to use histogram of oriented rectangles as the shape descriptor [7]. Likewise, Lin *et al.* [8] used silhouettes as the shape descriptor by counting the number of foreground pixels and motion-compensated optical flow as motion descriptor. Ikizler *et al.* [7] pointed out that human actions can be encoded as spatial information of body poses and dynamic information of body motions. As a matter of fact, some actions cannot be distinguished using shape or motion feature alone. For example, as shown in Fig. 1 a *skip* action may look very similar to a *run* action if only the pose of the body is observed. The classification task would be easier if the motion flow of the body is considered simultaneously. One would expect that *skip* action generates more vertical flows (upward/downward). Besides, actions such as jogging, walking and running can be easily confused if only pose information is used due to similar postures in the action sequences. Likewise, there are some actions which cannot be fully described by motion feature alone. Combining both cues potentially provides complementary information about an action. Conventionally, motion and shape feature vectors are concatenated to form a super vector [8,9]. However, the super vector may not explicitly convey the underlying action. Moreover, the super vector is unnecessarily long and require feature dimension reduction techniques. In this regard, an efficient representation of action is highly desirable. Motivated by the idea of convolution-based reverb in digital signal processing (DSP), we propose to encode the human action by convolving shape and motion histograms. This novel representation extracts rich information from actions.

Fig. 1. Similar poses observed in *skip* (left) and *run* (right) action sequences

The paper is organized as follows. In Section 2, we describe the details of shape and motion feature extraction. Next, we define the atomic action and explain how it can be represented as the convolved shape-motion histogram. Section 5 sets the backdrop for the experimental evaluation on Weizmann and KTH datasets while Section 6 shows the results. Finally, Section 7 gives the conclusion remarks of the paper.

2 Motion and Shape Histogram Binning

Observing shape and motion is a very natural way to recognize an action. Jhuang *et al.* [10] pointed out that the visual cortex in the brain has two pathways to process shape and motion information. Motivated by the robustness of the histogram of feature, we use histogram-of-oriented gradient (HOOG) and histogram-of-oriented optical flow (HOOF) as the shape and motion descriptors respectively. We adopt the histogram formation method which was originally introduced by Chaudhry *et al.* [11]. The method is more robust against scale variation and the change of motion direction. The method is illustrated in Fig. 4(a) with example of creating a 4-bin histogram. The main idea is to bin the vectors according to their primary angles from the horizontal axis. Therefore, the vectors are symmetry about the vertical axis. As a result, the histogram of a person moving from left to right will be same as the one with a person moving in the opposite direction. The contribution of each vector is proportional to its magnitude. The histogram is normalized to sum up to unity to make it scale-invariant. Therefore, we do not normalize the size of the bounding box. We further enhance Chaudhry *et al.* algorithm by including spatial information. This is done by dividing the bounding box of the subject into 4×4 regions as shown in Fig. 4(c).

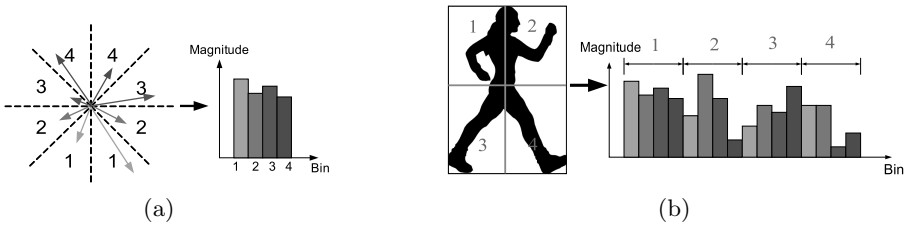


Fig. 2. (a) Histogram binning, (b) the bounding box is divided into 4×4 grid and the resultant histograms from each region are concatenated

3 Atomic Action Representation – Convolution of Shape-Motion Histrograms

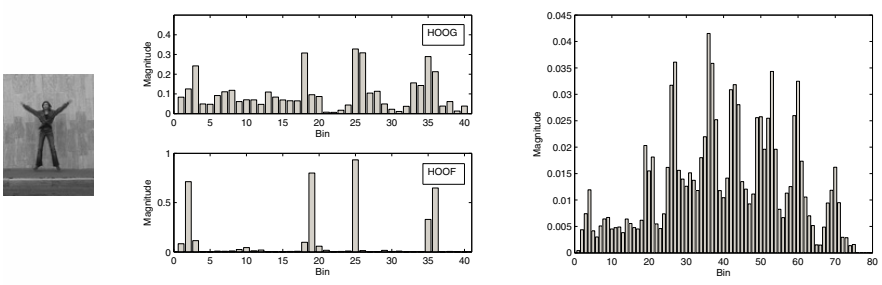
3.1 Defining 'Atomic Action'

Formally, a complex action can be decomposed to into a sequence of elementary building blocks, known as 'atomic actions'. For example, Fig. 3 shows a walking

action can be decomposed into several atomic actions – *right-leg stepping*, *two-leg crossing* and *left-leg crossing*. In this study, an atomic action is defined as the action performed at video frame t. It is represented by a shape histogram extracted at frame t and optical flow histogram computed at frames $(t-1)$ and t. Therefore, a T-frame action video has $(T-1)$ number of atomic actions.

Fig. 3. A walking sequence can be decomposed into sequence of atomic actions: *right-leg stepping* (left most), *two-leg crossing* (middle two), and *left-leg crossing* (right most)

3.2 Convolving Shape-Motion Histograms



(a) (b) (c)

Fig. 4. (a) A jumping-jack action, (b) the corresponding shape histogram (upper) and motion histogram (lower), (c) the convolved histogram

Inspired by the idea of convolution-based reverb in digital signal processing (DSP), we propose to encode atomic actions by convolving shape and motion histograms. In DSP, convolution is a mathematical way of combining two source signals to form an output signal. The output signal bears the characteristics of both sources. One important application of convolution is convolution-based reverb, a process for digitally simulating the reverberation of a virtual or physical space. Given the impulse response of a space which can be obtained by recording a short burst of a broad-band signal, we can convolve any "dry" signal (little room or space influence) with that impulse response. The results is that the sound appears to have been recorded in that space. Analogously, knowing that an action is characterized by both shape and motion information, we can

obtain an atomic action histogram $A[i]$ by convolving the corresponding shape histogram $X_s[i]$ and motion histogram $X_m[i]$:

$$A[i] = X_s[i] * X_m[i] = \sum_{k=-\infty}^{k=+\infty} X_s[k].X_m[i-k] \tag{1}$$

where the asterisk '*' denotes the convolution operator and square bracket [ ] indicates the signal is discrete. Since histograms are discrete in nature, for notation consistency the square bracket is omitted in the latter parts of this paper (i.e., x_i is equivalent to $x[i]$). Convolution operation is commutative meaning that it does not mathematically matter the order of the inputs. Thus, convolving X_s with X_m opposed X_m with X_s does not affect the result of the output. The length of output is given by the $\|X_s\| + \|X_m\| - 1$. Fig. 4 shows the convolution process of a jumping-jack action.

The proposed representation has two major advantages: First, the action histogram is more robust against noises. This is because each bin in the action histogram is influenced by bins in the shape histogram weighted by the motion histogram or vice versa (commutative property of convolution), therefore the effect of abrupt changes in the histogram magnitude can be minimized; second, the action histogram produced using convolution is more discriminative. We measure the ratio of inter-class distance to intra-class distance and the results on Weizmann dataset is shown in Table 1. We use Hellinger distance measure to compare two histograms:

$$D_h(X_1, X_2) = \left[1 - \sum_{\forall x} \sqrt{X_1 X_2}\right]^{\frac{1}{2}} \tag{2}$$

The results suggest that convolution operation produces the feature vectors that are potentially more discriminative than the features obtained through other combination methods.

4 Compact Video Representation: Distance Weighted Bag-of-Atomic-Actions

Over the past decade, a large body of work on human action recognition using local representation has been focusing on the bag-of-visual-words model [2, 3, 12–14]. In local representation, a huge collection of independent patches (e.g.: spatial-temporal features) is extracted from the training data. A codebook is then created from these local patches using some clustering algorithms such as K-means. Following that, an image or short sequence of images can be represented as a histogram which corresponds to the frequency of the visual-words. Apparently, this technique may not be directly applicable to global representation. Therefore, we propose an extension scheme to encode a video that is represented by the global features. In the proposed scheme, an action video is represented as a collection of repetitive atomic actions. Recall that an atomic

actions is fully characterized by the convolved shape-motion histogram as described in the previous section. A visual codebook can be created by performing K-means clustering on all atomic actions from the training data. The cluster centroids, which are essentially some normalized histograms, serve as the visual codewords. Next, each atomic action in the video is compared against those codewords and the distances are recorded accordingly. The distance between the atomic action and its nearest codeword is used to weight the histogram bin. The histogram for all relevant codewords in a video is computed by aggregating their respective distances. This final representation allows any lengthy video to be 'compressed' into a compact histogram. The histogram is normalized such that sum of the bins is unity. The normalization ensures that the histogram distribution is invariant to the video length. For instant, given a particular action class, we expect to see the codewords (i.e., key atomic actions) frequencies for a variable length video remains relatively stable.

Table 1. Comparison of normalized inter-intra class distance ratio on Weizmann dataset for different types of feature combination methods. Higher value indicates that the feature is potentially more discriminative.

Combination Strategies	Ratio
Convolution (*Conv*)	1.0000
Summation (*Sum*)	0.8535
Product (*Prod*)	0.8489
Concatenation (*Concat*)	0.8743

5 Experiments

We performed various experiments to evaluate the proposed action recognition framework on two publicly available datasets (see Fig. 5):

- **Weizmann.** The dataset was originally introduced in [5]. The dataset contains 90 low-resolution (180×144 pixels) video sequences with 9 subjects performing 10 actions: bend (*bend*), jumping-jack (*jack*), jump-forward (*jump*), jump-in-place (*pjump*), run (*run*), gallop-sideways (*side*), jump-forward-one-leg (*skip*), walk (*walk*), wave-one-hand (*wave1*), and wave-two-hands (*wave2*)[1]. We used the silhouettes provided to compute the bounding boxes for the subjects. HOOG and HOOF features are extracted from the silhouettes.

- **KTH.** The dataset was introduced in [15]. There are 25 subjects performing 6 actions: boxing, handclapping, handwaving, jogging, running, and walking. The low resolution (160×120) videos were recorded under four scenarios (s1-outdoors, s2- outdoors with scale variation, s3- outdoor with different clothes,

[1] Note that there are two versions of Weizmann dataset, the original one has 9 actions while the augmented version has 10 which includes *skip* action.

s4- indoors with lighting variation) and each video was split into 4 sub-clips. Originally, the dataset has (4 settings) × (25 subjects) × (6 actions) × (4 sub-clips) = 2400 clips. However, only 2391 clips are available because 8 clips were missing. We used the bounding box provided by Lin *et al.* [8] to locate the subject. Nevertheless, we did not compute the silhouette because object segmentation is not our focus in this work. Therefore, HOOG and HOOF features are extracted directly from the raw grayscale video frames.

In the literature, KTH dataset has been regarded either as one large set with strong intra-subject variations (*all-in-one*) or as four independent scenarios. In the latter case, each scenario is trained and tested separately. In this work, we only focus on the *all-in-one* case.

Leave-one-out cross validation (LOOCV) protocol is used in all evaluations. We use multiclass Support Vector Machine (SVM) as the classifier.

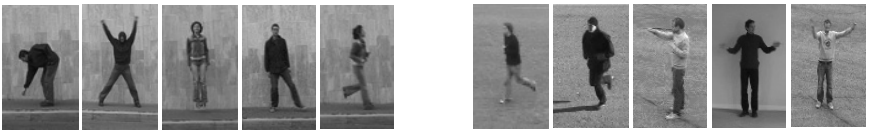


Fig. 5. Examples of different actions from databases Weizmann (left) and KTH (right)

6 Results

Table 2 shows the LOOCV recognition rate for Weizmann dataset. With only using 5 clusters (codewords), the convolved feature yielded a much higher accuracy (96.67%) compared to other features. When the number of clusters is increased further, the convolved feature consistently gives perfect classification accuracy (100%). It is worth noting that using only shape feature (HOOG) or motion feature (HOOF) resulting poorer results. On average, by using the proposed method we gain about 11.29% overall improvement (4% for *sum*, 5.33% for *prod*, 4.44% for *concat*, 23.56% for *HOOG* and 19.11% for *HOOF*). Since KTH is a more challenging dataset with strong intra-class variations, it would be interesting to find out if the proposed method can still perform well as for the Weizmann dataset. The results for KTH dataset is tabulated in Table 3. Obviously, similar to Weizmann dataset, for all number of clusters, higher accuracies are attained from the convolved feature. Although classification task on KTH dataset is more difficult, the advantage of using the convolved feature is more prominent on this dataset. The average improvement over all other five features is 19.56% (10.61% for *sum*, 10.58% for *prod*, 6.49% for *concat*, 42.72% for *HOOG* and 27.41% for *HOOF*). On this dataset, both HOOG and HOOF features fail to provide discriminative information which contribute to the poor classification results. One important observation from the results is that our method consistently requires a much smaller number of clusters (codewords) to give higher accuracy.

Table 2. Weizmann dataset: LOOCV classification accuracy using different number of clusters

No. of Clusters \Features	Convol	Sum	Prod	Concat	HOOG	HOOF
5	96.67	87.78	87.78	86.67	66.67	73.33
10	**100**	94.44	94.44	93.33	74.44	78.89
15	**100**	97.78	94.44	98.89	77.78	81.11
20	**100**	97.78	96.67	97.78	78.89	84.44
25	**100**	98.89	96.67	97.78	81.11	83.33

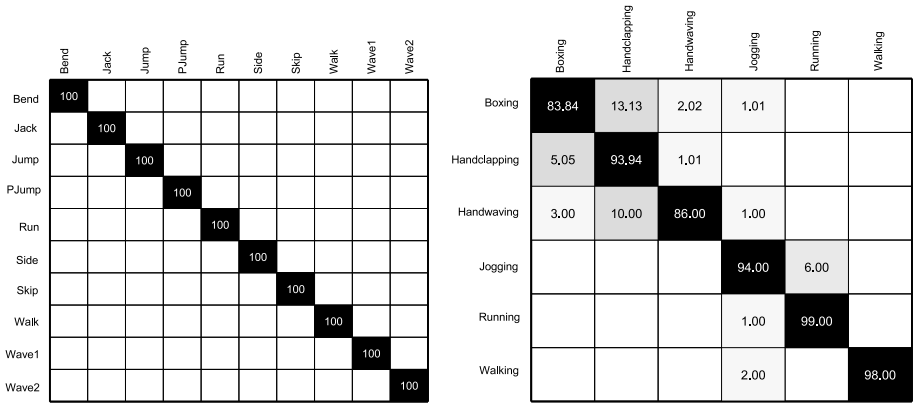
Table 3. KTH dataset: LOOCV classification accuracy using different number of clusters

No. of Clusters \ Features	Convol	Sum	Prod	Concat	HOOG	HOOF
10	83.94	70.25	72.22	75.58	45.88	57.73
25	91.63	79.94	79.92	83.30	51.90	63.88
40	**92.46**	82.44	81.43	87.64	45.24	64.88
55	91.46	84.43	83.62	86.97	45.58	63.37

For example, with only 10 clusters our method achieves comparable accuracy with the product feature which uses 40 clusters. This confirms the finding that the convolved feature is significantly more discriminative. Moreover, we notice that the product-based combination method is slightly inferior to the sum-based approach, probably due to the information lost when multiplying two features with elements equal to zero. The bottom part of Fig. 4(b) shows an optical flow histogram with many bins equal to zero. When multiplying this histogram with the gradient histogram, the output will bias to the motion input while some useful information from the shape input is discarded.

The confusion matrix for Weizmann and KTH datasets are given in Fig. 6(a) and 6(b) respectively. For KTH dataset, errors mainly occur when classifying *boxing*, *handwaving* and *handclapping*. The misclassification of boxing action could be due to the erroneous bounding box extracted which is off-centered from the body axis when the person punches to the side. Nevertheless, our method is able to discriminate *jogging-running-walking* classes very well. This is remarkable given that these three actions share substantial similarity in motion and shape cues.

We have compared the results with state-of-the-art action recognition approaches. Table 4 shows that our method achieved the same perfect accuracy as [9], [16] for Weizmann dataset. As for KTH dataset, it may be argued that the results are not directly comparable as different authors employed different evaluation protocol (splits vs. LOOCV). While not definitive, Table 5 still provides some indicative comparison. The result shows that our method achieves one of the highest accuracies (only slightly lower than [8]) as far as the LOOCV protocol is concerned. One possible reason is that Lin *et al.* used silhouette for feature



(a) (b)

Fig. 6. Confusion matrix for (a) Weizmann (accuracy = 100%), and (b) KTH dataset (accuracy = 92.46%)

Table 4. Comparison of Recognition Rates for Weizmann Dataset

Method	Accuracy (%)
Our method	**100.00**
Fathi [16]	100.00
Schindler [9]	100.00
Blank [5]	99.64
Jhuang [10]	98.80
Wang [17]	97.78
Chaudhry [11]	94.44
Niebles [14]	90.00

extraction while in our KTH experiment we only used the original grayscale image containing inside the bounding box. It is well known that silhouette-based approach is more robust but it requires good background modelling which is more restrictive than the bounding box-based approach. From the result, it can be deduced that our method can perform very well even without using silhouette. One biggest advantage of our approach is the simplicity in implementation as opposed to their complicated prototype trees generation procedure. Moreover, the feature used in our method has much lower dimension (length = 79) than those used in [8](length = 512) and [9](length = 1000). Most importantly, the tedious task of determining the optimal weight to control the relative importance of shape and motion cues during concatenation, is no longer required.

Table 5. Comparison of Recognition Rates for KTH Dataset

Method	Protocol	Accuracy (%)
Our method	LOOCV	**92.46**
Lin [8]	LOOCV	93.43
Schindler [9]	Splits	92.70
Fathi [16]	Splits	90.50
Ahmad [18]	Splits	88.83
Willems [3]	Splits	84.26
Niebles [14]	LOOCV	83.33
Dollár [2]	LOOCV	81.17
Ke [19]	LOOCV	80.90
Schüldt [15]	Splits	71.72

7 Conclusion

This paper presents a novel method to encode human actions by convolving shape-motion histograms. The inspiration comes from the success of convolution reverb application in digital signal processing. The main idea is to produce an output signal (i.e., action histogram) from the source signals (i.e., shape and motion histograms) so that the output shares the characteristics of both sources. The experimental results demonstrate that the proposed method is very efficient compared to other combination strategies such as sum, product and concatenation. Moreover, our results are compared to the state-of-the-art results.

References

1. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, Nice, France, pp. 432–439 (2003)
2. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, Beijing, China, pp. 65–72 (2005)
3. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
4. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV, Nice, France, pp. 726–733 (2003)
5. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, Beijing, China, pp. 1395–1402 (2005)
6. Ikizler, N., Cinbis, R.G., Duygulu, P.: Human action recognition with line and flow histograms. In: ICPR, Tampa, FL, pp. 1–4 (2008)
7. Ikizler, N., Duygulu, P.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image and Vision Computing 27, 1515–1526 (2009)
8. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: ICCV, Kyoto, Japan (2009)
9. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: CVPR, Anchorage, Alaska, pp. 1–8 (2008)

10. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV, Rio de Janeiro, Brazil, pp. 1–8 (2007)
11. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: CVPR, Miami, FL, USA, pp. 1932–1939 (2009)
12. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28, 976–990 (2010)
13. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. British Machine Vision Conference (BMVC), 127 (2009)
14. Niebles, J.C., Wang, H., Fei-fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Int'l J. Computer Vision 79, 299–318 (2008)
15. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR, Cambridge, United Kingdom, pp. 32–36 (2004)
16. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: CVPR, Anchorage, Alaska, pp. 1–8 (2008)
17. Wang, L., Suter, D.: Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: CVPR, Minnesota, USA, pp. 1–8 (2007)
18. Ahmad, M., Lee, S.: Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recognition 41, 2237–2252 (2008)
19. Ke, Y., Sukthankar, R., Hebert, M.: Spatio-temporal shape and flow correlation for action recognition. In: 7th Int. Workshop on Visual Surveillance (2007)