Recognition of Human Movements Using Hidden Markov Models - An Application to Visual Speech Recognition

Wai Chee Yau[1], Dinesh Kant Kumar[1] and Hans Weghorn[2]

[1] School of Electrical and Computer Engineering, RMIT University
GPO Box 2476V, Melbourne, Victoria 3001, Australia

[2] Information Technology, BA-University of Cooperative Education
Stuttgart, Germany

Abstract. This paper presents a novel approach for recognition of lower facial movements using motion features and hidden Markov models (HMM) for visual speech recognition applications. The proposed technique recognizes utterances based on mouth videos without using the acoustic signals. This paper adopts a visual speech model that divides utterances into sequences of smallest, visually distinguishable units known as visemes. The proposed technique uses the viseme model of Moving Picture Experts Group 4 (MPEG-4) standard. The facial movements in the video data are represented using 2D spatial-temporal templates (STT). The proposed technique combines discrete stationary wavelet transform (SWT) and Zernike moments to extract rotation invariant features from the STTs. Continuous HMM are used as speech classifier to model the English visemes. The preliminary results demonstrate that the proposed technique is suitable for classification of visemes with a good accuracy.

1 Introduction

Machine analysis of human motion is a growing research area with vast potential applications. The classification of human movements is challenging due to complex motion patterns of the human body [17]. Human movements are very diversified ranging from running to more subtle motions such as the facial movements while speaking. Identification of facial movements is an important aspect in recognition of utterances. Speech-based systems are emerging as attractive interfaces that provide the flexibility for users to control machines using speech.

In spite of the advancements in speech technology, speech recognition systems has yet to be used as mainstream human-computer interfaces (HCI). The difficulty of audio speech recognition systems is the sensitivity of such systems to changes in acoustic conditions. The performance of such systems degrades drastically when the acoustic signal strength is low, or in situations with high ambient noise levels [21]. To overcome this limitation, the non-acoustic modalities are used. Possible methods are such as visual [21], recording of vocal cords movements [5] and recording of facial muscle activity

Chee Yau W., Kant Kumar D. and Weghorn H. (2007).
Recognition of Human Movements Using Hidden Markov Models - An Application to Visual Speech Recognition.
In *Proceedings of the 7th International Workshop on Pattern Recognition in Information Systems*, pages 151-160
DOI: 10.5220/0002424901510160

[2]. The vision-based techniques are more desirable options as such techniques are non intrusive and do not require placement of sensors on the speaker.

Research where audio visual speech recognition (AVSR) systems are being made more robust, and able to recognize complex speech patterns are being reported [21, 10]. While AVSR systems are useful for applications such as for telephony in noisy environment, such systems are not suitable for people with speech impairment. AVSR systems are also not useful when it is essential to maintain silence. The need for visual-only, voice-less communication systems arises. Such systems are also known as visual speech recognition (VSR) systems.

The design of a typical pattern recognition system would involve three stages : 1) data acquisition and preprocessing, 2)data representation and, 3) decision making [11]. Similarly, the design of a VSR system consists of the recording and preprocessing of video, extraction of visual speech features and a speech classifier. The visual speech features represent the movements of the speech articulators such as the lips and jaw. The advantages of VSR system are : (i) not affected by audio noise (ii) not affected by change in acoustic conditions (iii) does not require the user to make a sound. The visual cues contain far less classification power for speech compared to audio data [21] and hence it is to be expected that VSR systems would have a small vocabulary.

Visual features proposed in the literature can be categorized into shape-based, pixel-based and motion-based features. The shape-based features rely on the shape of the mouth. The first VSR system was developed by Petajan [20] using shape-based features such as height and width of the mouth. Researchers have reported on the use of artificial markers on speaker's face to extract the lip contours [12, 1]. The use of artificial markers is not suitable for practical speech-based HCI applications. VSR systems that use pixel-based features assume that the pixel values around the mouth area contain salient speech information [14, 21].

Pixel-based and shape-based features extracted from static frames and can be viewed as static features. Such features attempt to model visual speech through the different static poses of the mouth in frames of the video. Features that directly utilize the dynamics of speech are the motion-based features. Few researchers have focused on motion-based features for VSR. The dynamics of the visual speech is important in the design and selection of visual features [22]. Goldschen et. al. [8] demonstrates that dynamic visual features are most discriminative when comparing static and motion features. One of the early motion features are based on the optical flow analysis [19]. Image subtraction techniques are used to extract motion-based features for visual speech recognition in [24]. Motion feature based on image subtraction are demonstrated to outperform optical flow analysis method [9]. This paper proposes a novel VSR technique based on motion features extracted using spatial-temporal templates (STT) to represent the dynamics of visual speech. STT are grayscale images that contain both spatial and temporal information of the motion in the video data. This paper proposes a system where the camera is attached in place of the microphone to the commonly available head-sets. The advantage of this is that it is no longer required to identify the region of interest, reducing the computation required. This paper reports on the use of wavelet transform and Zernike moments to extract rotation invariant features from the STT and hidden Markov models (HMM) to classify the features.

2 Background

2.1 Visual Speech Model

Speech can be organized as sequences of contiguous speech sounds known as phonemes. Visemes are the atomic units of visual movements associated with phonemes. This paper proposes the use of visemes to model visual speech. The motivation of using viseme as the recognition unit is because visemes can be concatenated to form words and sentences, thus providing the flexibility to increase the vocabulary of the system. The total number of visemes is much less than phonemes as speech is only partially visible [10]. While the video of the speaker's face shows the movement of the lips and jaw, the movements of other articulators such as tongue and glottis are often not visible. The articulation of different speech sounds (such as /p/ and /b/) may be associated with identical facial movements. Each viseme may corresponds to more than one phoneme, resulting in a many-to-one mapping of phonemes-to-visemes. It is difficult to differentiate phonemes with identical facial motions based solely on the visual speech signals and hence other information from other sensory components is required to disambiguate these phonemes.

There is no definite consensus about how the sets of visemes in English is constituted [10]. The number of visemes for English varies depending on factors such as the geographical location, culture, education background and age of the speaker. The geographic differences in English is most obvious where the sets of phonemes and visemes changes for different countries and even for areas within the same country. This paper adopts a viseme model established for facial animation applications by an international audiovisual object-based video representation standard known as MPEG-4. This model is selected to enable the proposed VSR system to be easily coupled with any MPEG-4 supported facial animation or speech synthesis systems to form an interactive speech-based HCI. Based on the MPEG-4 viseme model shown in Table 1, the English phonemes can be grouped into 14 visemes.

2.2 Motion Segmentation

This proposed technique adopts a motion segmentation approach based on spatial-temporal templates (STT) to extract the lower facial movements from the video data. STT are grayscale images that show where and when facial movements occurs in the video [3]. The spatial information of the motion is encoded in the pixel coordinates of the STT whereas the temporal information of the facial movements are implicitly represented by the intensity values of the pixels that varies linearly with the recency of the lower facial motion [26].

STT are generated by using accumulative image difference approach. Image subtraction is applied on the video of the speaker by subtracting the intensity values between successive frames to generate the difference of frames (DOF). The DOFs are binarised using an optimum threshold value, a that is determined through experimentation. The delimiters for the start and stop of the motion are manually inserted into the image sequence of every articulation. The intensity value of the STT at pixel location

Table 1. Viseme model of the MPEG-4 standard for English phonemes.

Viseme Number	Corresponding Phonemes
1	p, b, **m**
2	f,**v**
3	**T** , D
4	**t** , d
5	k, **g**
6	**tS**, dZ, S
7	**s** , z
8	**n** , l
9	**r**
10	**A:**
11	**e**
12	**I**
13	**Q**
14	**U**

(x, y) of t^{th} frame is defined by

$$STT_t(x,y) = max \bigcup_{t=1}^{N-1} B_t(x,y) \times t \tag{1}$$

where N is the total number of frames of the mouth video. $B_t(x,y)$ represents the binarised version of the DOF of frame t. In Eq. 1, $B_t(x,y)$ is multiplied with a linear ramp of time to implicitly encode the temporal information of the motion into the STT. By computing the STT values for all the pixels coordinates (x,y) of the image sequence using Eq. 1 will produce a grayscale image (STT) where the brightness of the pixels indicates the recency of motion in the image sequence. Figure 1 illustrates the STTs of fourteen visemes used in the experiments.

This motivation of using STT to segment the facial movements is because of the ability of STT to remove static elements and preserve the short duration facial movements in the video data. The STT approach is computationally inexpensive. Also, STT is insensitive to the speaker's skin color due to the image subtraction process. The speed of phonation of the speaker might vary for each repetition of the same phone. The variation in the speed of utterance results in the variation of the overall duration and there maybe variations in the micro phases of the utterances. The details of such variations are difficult to model due to the large inter-experiment variations. This paper suggests a model to approximate such variations by normalizing the overall duration of the utterance. This is achieved by normalizing the intensity values of the STT to in between 0 and 1.

STT is a view sensitive motion representation technique. STT generated from the sequence of images is dependent on factors such as position, orientation and distance of the speaker's mouth from the camera. Also STT is affected by small variations of the mouth movements while articulating the same phone. This paper proposes the use

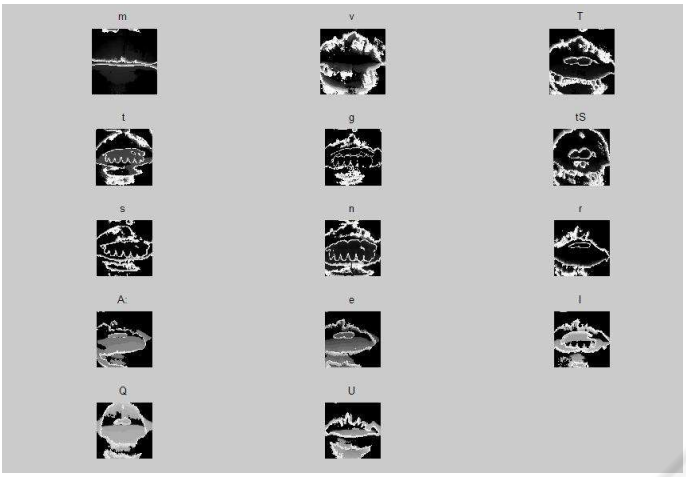

Fig. 1. Spatial-temporal templates (STT) of fourteen visemes based on the viseme model of MPEG-4 standard.

of approximate image of discrete stationary wavelet transform (SWT) to obtain a time-frequency representation of the STT that is insensitive to small variations of the facial movements. Figure 2 shows a block diagram of the proposed visual speech recognition technique.

2.3 Preprocessing of Spatial-Temporal Templates

This proposed technique uses discrete stationary wavelet transform (SWT) to obtain a transform representation of the STT that is insensitive to small variations of the facial movements. While the classical discrete wavelet transform (DWT) is suitable for this, DWT results in translation variance [16] where a small shift of the image in the space domain will yield very different wavelet coefficients. SWT restores the translation invariance of the signal by omitting the downsampling process of DWT, and results in redundancies. 2-D SWT at level 1 is applied on the STT to produce a spatial-frequency representation of the STT. Haar wavelet has been selected due to its spatial compactness and localization property. Another advantage is the low mathematical complexity of this wavelet. SWT decomposition of the STT generates four sub images. The approximate (LL) sub image is the smoothed version of the STT and carries the highest amount of information content among the four images. The LL subimage is used to represent the STT. The paper proposes to use Zernike moments to represent the SWT approximate image of the STT to reduce the dimension of the data.

2.4 Feature Extraction

Zernike moments are one of the image moments used in recognition of image patterns [13, 25]. Zernike moments have been demonstrated to outperformed other image moments such as geometric moments, Legendre moments and complex moments in terms

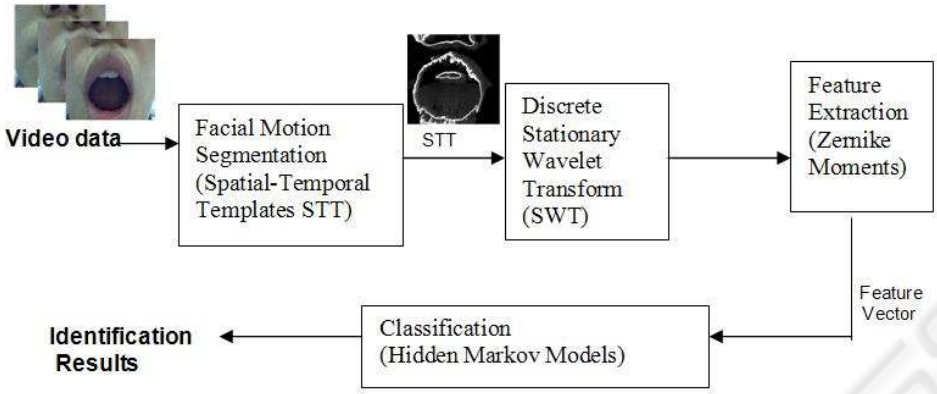

Fig. 2. Block Diagram of the Proposed Technique.

of sensitivity to image noise, information redundancy and capability for image representation [25]. The proposed technique uses Zernike moments as visual speech features to represent the SWT approximate image of the STT.

Zernike moments are computed by projecting the image function $f(x, y)$ onto the orthogonal Zernike polynomial V_{nl} of order n with repetition l is defined within a unit circle. The main advantage of Zernike moments is the simple rotational property of the features [13]. Zernike moments are also independent features due to the orthogonality of the Zernike polynomial V_{nl} [25]. Changes in the orientation of the mouth in the image result in a phase shift on the Zernike moments of the rotated image as compare to the original (non rotated image) [26]. Thus, the absolute value of Zernike moments are invariant to the rotation of the image patterns [13]. This paper uses the absolute value of the Zernike moments, $|Z'_{nl}|$ as the rotation invariant features. 49 Zernike moments that comprise of 0^{th} up to 12^{th} order moments are extracted from the approximate image of the STT.

2.5 Hidden Markov Models Classifier

To assign the motion features to an appropriate viseme group (class), a number of possible classifiers exists such as artificial neural network (ANN), support vector machines (SVM) and hidden Markov models (HMM). Left-right HMM is the most commonly used classifiers in speech recognition [21]. HMM is a finite state network that is constructed based on the theory of stochastic processes [6]. The strength of left-right HMM lies in its ability to statistically model the time-varying speech features [23].

A HMM is characterized by parameters the number of states in the model, the number of possible observation symbols, the state transition probability distribution ,observation symbol probability distribution and initial state distribution. Based on the appropriate values of HMM parameters, the HMM can generate a sequence of observation

vectors (feature vectors) by sampling a sequence of hidden states according to the transition probability distribution [23].

This paper adopts single-stream, continuous HMMs to classify the motion features of the visemes. Continuous HMMs is used as opposed to discrete HMMs to avoid the loss of information occur in the quantization of the features. The motion features are assumed to be Gaussian distributed. Each viseme is modelled using a left-right HMM with three states, one mixture of Gaussian component per state and diagonal covariance matrix. To use a continuous observation density in HMM, the observation probability distribution is assumed to be finite Gaussian mixture of the form

$$P(O_t|v) = \sum_{k=1}^{K_v} w_{v,k} N_l(O_t; m_{v,c}, s_{v,k}) \quad (2)$$

where the v represents the state of the HMM and O_t is the observed features at time t. The K_s mixture weights $w_{v,k}$ are positive and add to one, and $N_l(O; m, s)$ is the l-variate normal distribution with mean m and a diagonal covariance matrix s.

During training of the HMMs, the unknown HMMs parameters vectors consisting of the transition probability and observation probability are estimated iteratively using Expectation-Maximization (EM) algorithm [4] based on the training samples. In the classification stage, the unknown motion features are presented to the 14 trained HMMs and the features are assigned to the viseme class whose HMM produces output with the highest likelihood.

3 Methodology

Experiments were conducted to test the proposed visual speech recognition technique. The experiments were approved by the university's Human Experiments Ethics Committee. Fourteen visemes from the viseme model of MPEG-4 standard (highlighted in bold fonts in Table 1) were evaluated in the experiments. Video data was recorded using an inexpensive web camera in a typical office environment. This was done towards having a practical voiceless communication system using low resolution video recordings. The camera focused on the mouth region of the speaker and was kept stationary throughout the experiment. The following factors were kept the same during the recording of the videos : window size and view angle of the camera, background and illumination. 280 video files (240 x 240 pixels) were recorded and stored as true color (.AVI) files.The frame rate of the AVI files was 30 frames per second. One STT was generated from each AVI files. An example of STT for each visemes are shown in Figure 1. SWT at level-1 using Haar wavelet was applied on the STTs and the approximate image (LL) was used for analysis. 49 Zernike moments have been used as features to represent the SWT approximate image of the STT. The Zernike moments features were used to train the hidden Markov models (HMM) classifier. One HMM was created and trained for each viseme. The leave-one-out method was used in the experiment to evaluate the performance of the proposed approach. The HMMs were trained with 266 training samples and were evaluated on the 14 remaining samples (1 sample from each viseme group). This process is repeated 20 times with different sets of training and testing data. The average recognition rates of the HMMs for the 20 repetitions were computed.

4 Results and Discussion

The classification accuracies of the HMM are tabulated in Table 2. The average recognition rate of the proposed visual speech recognition system is 88.2%. The results indicate that the proposed technique based on motion features is suitable for viseme recognition.

Table 2. Recognition Rates of the proposed system based on viseme model of MPEG-4 standard.

Viseme	Recognition Rate (%)
m	95
v	90
T	70
t	80
g	85
tS	95
s	95
n	40
r	100
A:	100
e	100
I	95
Q	95
U	95

Based on the results, the proposed technique is highly accurate for vowels classification using the motion features. An average success rate of 97% is achieved in recognizing vowels. The classification accuracies of consonants are slightly lower due to the poor recognition rate of one of the consonant - /n/. One of the possible reason for the misclassifications of /n/ is due to the inability of vision-based technique to capture the occluded speech articulators movements. The movement of the tongue within the mouth cavity is not visible (occluded by the teeth) in the video data during the pronunciation of /n/. Thus, STT of /n/ does not contain information on the tongue movement which may have resulted a high error rate for /n/.

To compare the results of the proposed approach with other related work is inappropriate due to the different video corpus and recognition tasks used. In a similar visual-only speech recognition task (based on the the 14 visemes of MPEG-4 standard) reported in [7], a similar error rate was obtained using shape-based features (geometric measures of the lip) extracted from static images. Nevertheless, the errors made in our proposed system using motion features are different compare to the errors reported in [7] that uses static features. This indicates that complementary information exist in static and dynamic features of visual speech. For example, our proposed system has a much lower error rate in identifying visemes /m/, /t/ and /r/ by using the facial movement features as compare to the results in [7]. This shows that motion features

are better in representing phones which involve distinct facial movements (such as the bilabial movements of /m/). The static features of [7] yield better results in classifying visemes with ambiguous or occluded motion of the speech articulators such as /n/.

The results demonstrate that a computationally inexpensive system which can easily be developed on a DSP chip for voice-less(mime speech) communication application. The proposed system has been designed for specific applications such as control of machines using simple commands consisting of discrete utterances without requiring the user to make a sound.

5 Conclusion

This paper reports on a facial movement classification technique using HMM for visual speech recognition. The proposed technique recognizes utterances based on the visible movements of the jaw and mouth of the speaker. The facial movements of the video data are represented using spatial-temporal templates (STT). This paper adopts the MPEG-4 viseme model as the visual model to represent English phonemes. The proposed technique employs continuous hidden Markov models (HMM) as the supervised classifier. A low error rate of 12% is obtained in classifying the visemes using the proposed approach. Our results demonstrate that the proposed technique based on motion features is suitable for facial movement recognition.

The differences in classification errors of the proposed technique using motion features compare to the approach using static features [7] suggest that complementary information may exists between dynamic and static features. For future work, the authors intend to combine static and dynamic features for recognition of facial movements.

Such a system could be used to drive computerized machinery in noisy environments. The system may also be used for helping disabled people to use a computer and for voice-less communication.

References

1. Adjoudani, A., Benoit, C., Levine, E.P.: On the Integration of Auditory and Visual Parameters in an HMM-based ASR. Speechreading by Humans and Machines: Models, Systems, and Applications, Springer,(1996) 461–472
2. Arjunan, S. P., Kumar, D. K., Yau, W. C., Weghorn, H. : Unspoken Vowel Recognition Using Facial Electromyogram. IEEE EMBC, New York, (2006)
3. Bobick, A. F., Davis, J. W.: The Recognition of Human Movement Using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23 (2001) 257–267
4. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royyal Statist. Soc., Vol. 39 (1977) 1–38
5. Dikshit, P. S., Schubert, R. W.: Electroglottograph as an additional source of information in isolated word recognition. Fourteenth Southern Biomedical Engineering Conference (1995) 1–4
6. Fred, A., Marques, J. S., Jorge, P. M. : Hidden Markov models vs syntactic modeling in object recognition. In Intl Conference on Image Processing, ICIP97, Santa Barbara (1997) 893-896

7. Foo, S. W., Dong, L.: Recognition of Visual Speech Elements Using Hidden Markov Models. Lecture notes in computer science, Springer-Verlag, Vol. 2532 (2002) 607–614
8. Goldschen, A. J., Garcia, O. N., Petajan, E.:Continuous Optical Automatic Speech Recognition by Lipreading.presented at 28th Annual Asilomar Conf on Signal Systems and Computer(1994)
9. Gray, M. S., Movella, J. R., Sejnowski, T. J.: Dynamic features for visual speechreading : a systematic comparison. 3rd Joint Symposium on Neural Computation, La Jolla (1997)
10. Hazen, T. J.: Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition. IEEE Transactions on Audio, Speech and Language Processing (2006) Vol. 14 No. 3 1082–1089
11. Jain, A. K., Duin, R. P. W., Mao, J. : Statistical Pattern Recognition : A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1 (2000) 4–37
12. Kaynak, M. N., Qi, Z., Cheok, A. D., Sengupta, K., Chung, K. C. : Audio-visual modeling for bimodal speech recognition. IEEE Transactions on Systems, Man and Cybernetics, (2001) Vol. 34 564–570
13. Khontazad, A., Hong, Y. H.: Invariant Image Recognition by Zernike Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence (1990) Vol. 12 489–497
14. Liang, L., Liu, X. , Zhao, Y., Pi, X., Nefian, A. V.: Speaker Independent Audio-Visual Continuous Speech Recognition. In IEEE Int. Conf. on Multimedia and Expo (2002)
15. Lippmann, R. P.: Speech recognition by machines and humans. J. Speech Communication (1997) Vol. 22 1–15
16. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1998)
17. Moeslund, T. B., Hilton, A., Kruger, V. : A survey of advances in vison-based human motion capture and analysis. Computer Vision and Image Understanding, Vol. 104 (2006) 90–126
18. Manabe, H., Hiraiwa, A. : Unvoiced speech recognition using EMG - mime speech recognition. Conference on Human Factors in Computing Systems CHI '03,Ft. Lauderdale, Florida, USA, (2003) 794–795
19. Mase, K., Pentland, A.: Automatic lipreading by optical-flow analysis. Systems and Computers in Japan, (1991) Vol. 22, 67–76
20. Petajan, E. D.: Automatic Lip-reading to Enhance Speech Recognition. In GLOBECOM'84,IEEE Global Telecommunication Conference (1984)
21. Potamianos, G., Neti, C., Gravier, G., Senior, A. W.: Recent Advances in Automatic Recognition of Audio-Visual Speech. In Proc. of IEEE, Vol. 91 (2003)
22. Potamianos, G., Neti, C., Luettin, J., Matthews, I.: Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing, (2004)
23. Rabiner, L. R. : A tutorial on HMM and selected applications in speech recognition.Proc. IEEE, Vol. 77, No. 2, Issue 2 ,(1989) 257–286
24. Scanlon, P., Reilly, R. B., Chazal, P.D.: Visual Feature Analysis for Automatic Speechreading. Proceedings of Audio Visual Speech Processing Conf., France, (2003)
25. Teh, C. H., Chin, R. T.: On Image Analysis by the Methods of Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 10. (1988)496–513
26. Yau, W. C., Kumar, D. K., Arjunan, S. P. : Visual Speech Recognition Method Using Translation, Scale and Rotation Invariant Features. IEEE International Conference on Advanced Video and Signal based Surveillance, Sydney, Australia (2006)