# Patch-Based Analysis of Visual Speech from Multiple Views

*Patrick Lucey*[1], *Gerasimos Potamianos*[2], *Sridha Sridharan*[1]

[1]Speech, Audio, Image and Video Technology Laboratory,
Queensland University of Technology, Brisbane, QLD, 4000, Australia
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

p.lucey@qut.edu.au, gpotam@us.ibm.com, s.sridharan@qut.edu.au

## Abstract

Obtaining a robust feature representation of visual speech is of crucial importance in the design of audio-visual automatic speech recognition systems. In the literature, when visual appearance based features are employed for this purpose, they are typically extracted using a "holistic" approach. Namely, a transformation of the pixel values of the entire region-of-interest (ROI) is obtained, with the ROI covering the speaker's mouth and often surrounding facial area. In this paper, we instead consider a "patch" based visual feature extraction approach, within the appearance based framework. In particular, we conduct a novel analysis to determine which areas (patches) of the mouth ROI are the most informative for visual speech. Furthermore, we extend this analysis beyond the traditional frontal views, by investigating profile views as well. Not surprisingly, and for both frontal and profile views, we conclude that the central mouth patches are the most informative, but less so than the holistic features of the entire ROI. Nevertheless, fusion of holistic and the best patch based features further improves visual speech recognition performance, compared to either feature set alone. Finally, we discuss scenarios where the patch based approach may be preferable to holistic features.

**Index Terms**: Audio-visual automatic speech recognition (AVASR), multi-view, lipreading, visual features, patches.

## 1. Introduction

There has been significant interest and research work over the past few years on the subject of audio-visual automatic speech recognition (AVASR), due to the benefits of visual speech information to ASR robustness to noise. Few highlights of such work include large-vocabulary, speaker-independent AVASR [1], experiments on realistic audio-visual environments, such as offices and automobiles [2, 3], design of a wearable audio-visual headset to robustly capture the speaker's mouth [4], real-time AVASR algorithmic implementation into a demoable system [5], and, most recently, AVASR from non-frontal (profile) views [6, 7].

Much of the extensive literature works on this subject emphasize the fact that obtaining a robust feature representation of visual speech is of crucial importance to the design of AVASR systems. Such features are most often based on visual appearance of the mouth region, although alternative approaches exist that employ shape based features or combinations of both [8]. In the popular appearance based feature extraction scheme, the features are obtained using a "*holistic*" approach: A transformation of the pixel values of the entire region-of-interest (ROI) is performed, with the ROI covering the speaker's mouth and often surrounding facial area, as in [9]. There, feature extraction consists of a cascade of linear transforms that captures both spatial and temporal visual speech components from a sequence of mouth ROIs; the first step in this cascade is a discrete cosine transform of the *entire* ROI. A potential problem with such holistic approach is that these features may not take into account all possible changes that occur within the mouth region during articulation (process of changing the shape of the vocal tract using the articulators, i.e., lips and jaw). Conversely, some features may be assigned ineffectively on relatively "unimportant" regions of the mouth. This is particularly undesirable in the statistical modeling process that follows feature extraction. This process typically employs a hidden Markov model (HMM) framework, which requires low-dimensionality vectors (normally less than 60) to ensure generalization and avoid the curse of dimensionality [10].

Motivated by the above, in this paper, we deviate from the holistic feature extraction paradigm, proposing instead a "*patch*" based visual feature extraction scheme, within the appearance based framework. In particular, we conduct a novel analysis to determine which areas (patches) of the mouth ROI are the most informative for visual speech. This is accomplished by essentially "breaking" the ROI up into an ensemble of image patches, subsequently modeling and recognizing visual speech from each patch individually. This approach could have a number of potential benefits: For example, if it is determined that there is a tendency for a particular area of the ROI to be more useful in terms of lipreading than others, that particular area could be weighted more to improve performance over the current holistic representation; in addition, this approach could be more robust to localized visual noise. The two feature extraction paradigms (holistic vs. patch based) are depicted in Fig. 1.

Patch-based analysis of the ROI is heavily motivated by work in face recognition. Techniques that decompose the face into an ensemble of salient patches have reported superior face recognition performance compared to approaches that treat the face as a whole [11, 12, 13]. The idea behind breaking the face into a series of patches is that it is easier to take into account the local changes in appearance due to the complicated three-dimensional facial shape, in comparison to treating it holistically [14]. Furthermore, as no similar prior work has been conducted in the area of AVASR, our proposed patch-based investigation could provide an understanding as to which areas of the ROI are the most pertinent to visual speech.

We conduct all experiments for this paper on both frontal and profile view data. For this purpose, we employ a suitable multi-view database, as described in Section 2. Furthermore, we concentrate entirely on the problem of auto-

Copyright © 2008 AVISA

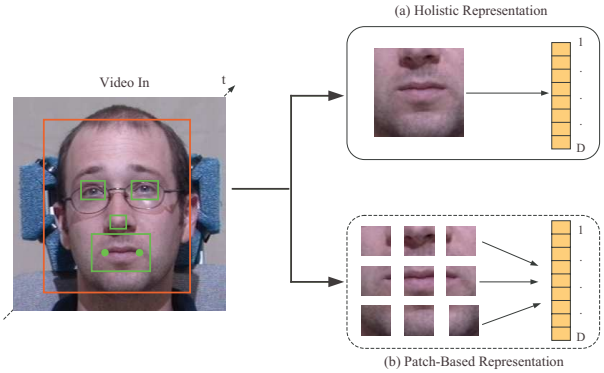


Figure 1: *Overview of the holistic and patch-based visual feature extraction approaches considered in this paper – depicted for the case of a frontal view frame. Following extraction of the mouth region-of-interest (ROI), the holistic approach (*top*) extracts appearance visual features (based on image transforms) of the entire ROI. Instead, the patch based approach (*bottom*) considers appearance based features extracted from each of nine patches separately. Such patch features could eventually be combined with the holistic ones (as described in our experiments – see Section 4), or even fused across patches into a single representation employing a multi-steam hidden Markov model of visual speech (future work).*

matic speechreading (visual-only ASR). Such focus prevents our comparative results from being skewed by the audio modality and the audio-visual fusion component used. The experiments are reported in Section 4, following a presentation of the lipreading system components in Section 3. Finally, Section 5 concludes the paper.

## 2. The IBM Smart-Room Database

As discussed in the Introduction, we are interested in applying our patch based feature representation idea on both frontal and profile view data. A suitable corpus for this purpose is the IBM smart room database collected as part of the recently concluded "Computers in the Human Interaction Loop" (CHIL) [15] integrated project, funded by the European Union.

The corpus contains a total of 38 subjects uttering connected-digit strings, using two microphones and three PTZ cameras. Of the two microphones, one is head-mounted (close-talking channel – see also Figure 2), and the other is omni-directional, located on a wall close to the recorded subject (far-field channel). The three PTZ cameras record frontal and two

Figure 2: *Examples of synchronous frontal and profile video frames of four subjects from the audio-visual database used in this paper.*

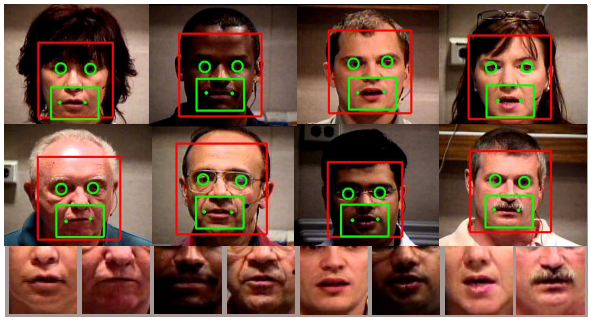

Figure 3: *Mouth ROI extraction examples for frontal views. The upper rows show examples of the localized face, eyes, mouth region, and mouth corners. The lower row depicts the corresponding normalized mouth ROIs of size 32×32 pixels.*

side views of the subject, and feed a single video channel into a laptop via a quad-splitter and an S-video–to–DV converter. As a result, two synchronous audio streams at 22kHz and three visual streams at 30 Hz and 368 × 240-pixel frames are available. Among these available streams, two video views are employed in this work, namely the frontal and right profile (which is the one "closest" to the profile pose – see Figure 2). A total of 1661 utterances are used in the experiments, partitioned using a multi-speaker paradigm into 1198 sequences for training (1 hr 51 min in duration), 242 for testing (23 min), and 221 sequences (15 min) that are allocated to a held-out set.

## 3. The Lipreading System

In this Section, we proceed to describe the basic components of the automatic speechreading (lipreading) system used in the paper, for both frontal and profile view data. In particular, we discuss ROI extraction, holistic and patch-based feature representation, concluding with an overview of the employed HMM-based statistical modeling of visual speech.

### 3.1. ROI Tracking for Frontal and Profile Views

For this paper, we use the AdaBoost framework of Viola and Jones [16], later extended by Leinhart and Maydt [17], to perform the mouth ROI localization and extraction. This framework allows us to generate face and facial feature localizers specific for each view-point, but nevertheless using a consistent approach across both views. These classifiers are trained using the OpenCV libraries [18], and their application requires that the speaker pose is first determined (an issue that is overlooked in this paper). Following this step, ROIs are obtained for each view at the same resolution (32×32 pixels), and visual feature vectors are extracted using the same approach for both views.

The actual task of mouth detection and ROI extraction was performed as follows: Given the video of a spoken utterance, the face detector of the specific pose was applied to estimate the location of the speaker's face. For the frontal scenario, once the face was found, the two eyes were detected and then a coarse mouth region was obtained. From this estimate, we applied detectors to find the corners of the mouth. From these detected lip corners, a normalized 32×32-pixel ROI was then extracted for use in our lipreading system. For the right profile case, once the face was found, the left eye and the nose were detected. From these located features, a coarse mouth detector was ap-

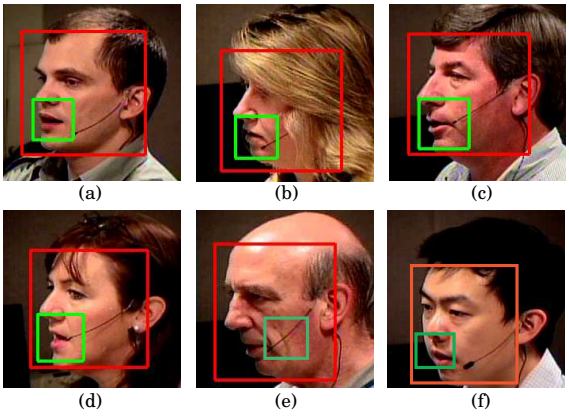


Figure 4: *Examples of accurate (a-d) and inaccurate (e,f) results of the profile-view localization and tracking system. In (f), it can be seen that the subject exhibits a somewhat more frontal pose compared to the profile view of the other subjects.*

plied to give an estimate of the mouth region. From there, we detected the mouth center and the left mouth corner. A normalized 32×32-pixel profile mouth ROI was then extracted, based on the distance from the left mouth corner to the left eye. These two points were used as reference points, as they were the most reliable to detect. More information can be found in [6]. As the Adaboost framework allows for extremely quick detection, we were able to perform detection on every frame and used median filtering to allow for smooth tracking. Examples for the frontal and profile extracted ROIs are given in Figs. 3 and 4, respectively.

### 3.2. Holistic Visual Feature Extraction

For both frontal and profile views, the same visual feature extraction process was applied. Following ROI extraction, the mean ROI over the utterance was removed. This approach is very similar to cepstral mean subtraction (CMS) in the audio domain and is known as feature mean normalization (FMN). Our implementation is similar to that of [8], however in our approach we performed normalization in the image domain instead of the feature domain. A two-dimensional, separable, discrete cosine transform (DCT) was then applied on the resulting mean-removed ROI, with 100 DCT coefficients retained, according to a zig-zag pattern. An intra-frame linear discriminant analysis (LDA) step was then used to project the features down to 30 dimensions, resulting in a "static" visual feature vector. Subsequently, in order to incorporate dynamic speech information, five of these neighboring static feature vectors over $\pm$ 2 adjacent frames were concatenated, and were projected via an inter-frame LDA step to yield a "dynamic" visual feature vector of dimension 40, extracted at the video frame rate of 30 Hz. The classes used for LDA matrix calculation were the HMM states (see Section 3.4), based on forced alignment employing an audio-only HMM on the far-field audio channel of the database.

### 3.3. Patch-Based Visual Feature Extraction

In contrast to the holistic approach, in the patch based system the ROI (frontal or profile) is decomposed into smaller regions. In this paper, we have chosen nine square patches of size 16×16 pixels each, with a 50% overlap with neighboring ones. Examples of these patches are depicted in Figs. 5 and 6 for the frontal and profile cases. The patches are numbered sequentially as shown in these figures. Notice that in both cases, patch number 5 contains most of the central mouth region information.

Following patch extraction, visual features are obtained in an identical fashion to the holistic approach. Namely, 100 DCT coefficients are retained for each 16×16-pixel patch, giving rise to 40-dimensional features per patch at 30 Hz, following the intra- and inter-frame LDA steps described in Section 3.2.

### 3.4. Visual Speech Modeling

Following the extraction of holistic or patch-based visual features, these can be fed into an automatic speechreading system to yield an estimate of the spoken word sequence. In this work, we employ an HMM based ASR system for this purpose. In particular, for the connected-digit recognition task considered here, eleven nine-state, left-to-right, whole-word models are used, one for each digit (both "oh" and "zero" are included), with seven Gaussian mixtures per state. A silence and short-pause model are also employed. All models are bootstraped from a segmentation of the audio channel of the database, obtained by an audio-only HMM with identical topology, and trained by the expectation-maximization algorithm. For testing, Viterbi decoding is used with no grammar or language model present (i.e., no constraints are imposed on the digit string length). The HTK toolkit is utilized for both system training and testing [19].

Such HMMs are trained on both holistic visual features, as well as for each of the patch based feature representations, since we are interested in comparing speechreading performance between the two approaches as well as across the various patches. In addition, in our experiments in Section 4, we also combine patch-based models with the holistic HMM. This is performed employing the decision fusion framework by means of a two-stream HMM [8]. In this approach, concatenated holistic and patch features are considered generated by the two-stream HMM, arising by combining two single-stream HMMs of identical topology (states and transitions), one modeling the holistic features, the other the patch based ones. The state-conditional observation log-likelihood of the resulting HMM is a linear combination of the ones of its two single-stream HMM components. In the experiments reported in Section 4, the HMM parameters are obtained using the expectation-maximization algorithm [19]. The weights employed in the linear combination of the two log-likelihoods are estimated at the end of the training procedure, by minimizing the word error rate on the held-out data set (see Section 2).

## 4. Experiments

Following the overview of the speechreading system components, we next proceed with our experiments. These are grouped into two subsections, one for each of the two views of interest.

### 4.1. Frontal-View Experiments

As already described in Section 3.3, for frontal views we consider nine 16×16-pixel patches as a decomposition of the frontal holistic ROI (see also Fig. 5). Following this step, 40-dimensional visual features are extracted, and HMMs are trained for each patch. Recognition results are depicted in Table 1 and are compared to the holistic system (40-dimensional visual features on the entire ROI).

These results suggest that most visual speech information

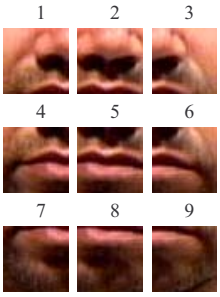


Figure 5: *Examples of the frontal-view ROI, decomposed into nine patches. The patches are numbered 1 to 9, from top-to-bottom, and left-to-right, as depicted in the figure.*

| | *WER (%)* | | |
|---|---|---|---|
| *Patches 1–3* | 47.53 | 54.80 | 49.19 |
| *Patches 4–6* | **33.98** | **33.94** | **33.46** |
| *Patches 7–9* | 39.92 | 38.55 | 47.86 |
| *Holistic* | 27.66 | | |

Table 1: *Frontal-view lipreading performance of each of the nine 16×16-pixel patch-based systems, also compared to the holistic approach. All results are in word error rate (WER), %.*

stems from the middle band of the ROI (patches 4–6). This of course is not surprising, as these ROI areas contain most visible articulators such as the lips, teeth and tongue. It can be seen that the area of the ROI that contains the least amount of visual speech information is patch 2, which corresponds to the nose and surrounding areas. This shows that the top of the ROI is the least effective for lipreading due to its fixed nature.

These results highlight a potential problem with the holistic approach. Noting that most of the lipreading performance stems from the ROI center (patches 4–6), it is a possibility that when executing the holistic approach, some of this speech discrimination power is diminished in an effort to incorporate the entire ROI into the feature representation. To investigate whether this is the case or not, we fuse the holistic representation with each 16×16 pixel patch. The hope is that any important information, possibly lost or diminished in the holistic representation, will be reenforced by the introduction of a local patch. In these experiments, only the holistic and individual patches are used, combined by means of a two-stream HMM. In particular, 40-dimensional holistic features and 20-dimensional patch-based ones are fused, in an effort to keep the concatenated feature dimensionality low. The results are reported in Table 2.

These results suggest that by fusing each patch with the

| | *WER (%)* | | |
|---|---|---|---|
| *Patches 1–3* | 27.70 | 27.98 | 27.67 |
| *Patches 4–6* | 26.84 | **26.76** | 26.79 |
| *Patches 7–9* | 27.02 | 27.15 | 28.21 |
| *Holistic* | 27.66 | | |

Table 2: *Frontal-view lipreading performance of each individual patch fused together with the holistic system by means of a two-stream HMM. The stand-alone holistic system performance is also depicted, for reference.*


Figure 6: *Examples of the profile-view ROI, decomposed into nine patches. The patches are numbered 1 to 9, from top-to-bottom, and left-to-right, as depicted in the figure.*

| | *WER (%)* | | |
|---|---|---|---|
| *Patches 1–3* | 69.94 | 64.97 | 61.93 |
| *Patches 4–6* | 55.32 | **48.48** | **49.38** |
| *Patches 7–9* | 58.60 | **49.67** | 66.49 |
| *Holistic* | 38.88 | | |

Table 3: *Profile-view lipreading performance of each of the nine 16×16-pixel patch-based systems, compared to the holistic approach.*

holistic representation, a slight improvement over the holistic-only result for most patches can be achieved (except for patch 2). This appears to support the hypothesis that some important visual speech classification information is lost, when visual features are calculated for the entire patch. However, by fusing the features of more salient regions with holistic ones, some of this important local information can be retained, thus improving overall lipreading performance. This is highlighted by the performance of patch 5 features, when fused with holistic ones, that achieves a 26.76% WER, as compared to 27.66% of the holistic representation alone. Nevertheless, this represents a rather small improvement at the price of a significant computational increase.

### 4.2. Profile-View Experiments

Similarly to Section 4.1, and as described in Section 3.3, for profile views we consider nine 16×16-pixel patches as a decomposition of the profile holistic ROI (see also Fig. 6). Following this step, 40-dimensional visual features are extracted and HMMs trained for each patch. Recognition results are depicted in Table 3, and are also compared to the holistic system (40-dimensional features on the entire profile ROI).

Not surprisingly, these results demonstrate that the region

| | *WER (%)* | | |
|---|---|---|---|
| *Patches 1–3* | 39.83 | 39.34 | 39.20 |
| *Patches 4–6* | 39.04 | **38.51** | 38.89 |
| *Patches 7–9* | 39.27 | 38.91 | 39.53 |
| *Holistic* | 38.88 | | |

Table 4: *Profile-view lipreading performance of each individual patch fused together with the holistic system employing a two-stream HMM. The stand-alone holistic system performance is also shown.*

containing the lips and jaw is the most useful for lipreading (patches 5, 6, and 8). This again backs up the hypothesis that movement of the visible articulators is of most benefit to recognizing visual speech. As for the frontal case, the nose region appears to be of little value for lipreading (patch 2), as well as the regions which contain the background (patches 1 and 7), or the skin around the lips (patches 3 and 9). Note however that background patches 1 and 7 may contain important lip protrusion information, possibly complementary to the frontal view.

To determine if any information in the holistic representation is lost by including the less pertinent areas of the profile ROI, fusion of each of the patches is performed with the holistic representation using a two-stream HMM. The results for these experiments are depicted in Table 4. Similarly to the frontal view, only a slight improvement over the holistic system is gained from fusing the middle patch (patch 5) – a WER of 38.71% compared to 38.88%. For all other patches, similar or worse performance is achieved, which suggests that little or no additional information is included by this approach.

## 5. Conclusions

In this paper we conducted a novel analysis using patches applied on both the frontal and profile mouth ROIs to determine the saliency of their various parts in the task of visual speech recognition. We showed that in both views, the middle patch containing most visible articulators, such as the lips, teeth, and tongue, provided the most visual speech information for automatic speechreading. However this information was less than that of holistic features extracted from the entire ROI. Nevertheless, fusion of holistic and the best patch based features slightly improved visual speech recognition performance, compared to the holistic approach, at an increased computational cost.

This work represents our first effort to deviate from the traditional holistic visual appearance feature extraction schemes, popular in the AVASR literature. In future work, we will investigate the possibility of fusing the patch-based features across the various patches, by employing an appropriate multi-stream HMM. This framework will allow allocating individual weights to the various patches, based on their contribution to overall lipreading performance. This approach is expected to potentially be of benefit in several scenarios, for example when localized visual noise corrupts specific patches, or when mouth ROI asymmetry is present.

## 6. Acknowledgements

The QUT portion of this research was supported by the Australian Research Council Grant No:LP0562101.

## 7. References

[1] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., & Vergyri, D., "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop," In *Proceedings of the Workshop on Multimedia Signal Processing*, (Cannes, France), 619–624, 2001.

[2] Potamianos, G. & Neti, C., "Audio-visual speech recognition in challenging environments," In *Proceedings of the European Conference on Speech Communication and Technology*, (Geneva, Switzerland), 1293–1296, 2003.

[3] Libal, V., Connell, J., Potamianos, G., & Marcheret, E., "An embedded system for in-vehicle visual speech activity detection," In *Proceedings of the International Workshop on Multimedia Signal Processing*, (Chania, Greece), 255–258, 2007.

[4] Huang, J., Potamianos, G., Connell, J., & Neti, C., "Audio-visual speech recognition using an infrared headset," *Speech Communication, 44*, 83–96, 2004.

[5] Connell, J., Haas, N., Marcheret, E., Neti, C., Potamianos, G., & Velipasalar, S., "A real-time prototype for small-vocabulary audio-visual ASR," In *Proceedings of the International Conference on Multimedia and Expo*, (Baltimore, MD, USA), 469–472, 2003.

[6] Lucey, P. & Potamianos, G., "Lipreading using profile versus frontal views," In *Proceedings of the International Workshop on Multimedia Signal Processing*, (Victoria, Canada), 24–28, 2006.

[7] Lucey, P., Potamianos, G., & Sridharan, S., "A unified approach to multi-pose audio-visual ASR," In *Proceedings of the Conference of the International Speech Communication Association*, (Antwerp, Belgium), 650–653, 2007.

[8] Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A.W., "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE, 91*(9), 1306–1326, 2003.

[9] Potamianos, G. & Neti, C., "Improved ROI and within frame discriminant features for lipreading," In *Proceedings of International Conference on Image Processing*, (Thessaloniki, Greece), 250–253, 2001.

[10] Bishop, C., *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] Brunelli, R. & Poggio, T., "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*, 1042–1052, 1993.

[12] Moghaddam, B. & Pentland, A., "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 696–710, 1997.

[13] Martinez, A., "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(6), 748–763, 2002.

[14] Lucey, S. & Chen, T., "Learning patch dependencies for improved pose mismatched face verification," In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, (New York, NY, USA), 909–915, 2006.

[15] The CHIL project: Computers in the Human Interaction Loop. [online] `http://chil.server.de`

[16] Viola, P. & Jones, M., "Rapid object detection using a boosted cascade of simple features," In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, (Kauai, HI, USA), 511–518, 2001.

[17] Leinhart, R. & Maydt, J., "An extended set of Haar-like features," In *Proceedings of the International Conference on Image Processing*, (Rochester, NY, USA), 900–903, 2002.

[18] OpenCV: Open Source Computer Vision Library. [online] `http://sourceforge.net/projects/ovencvlibrary`

[19] Young, S., Everman, G., Hain, T., Kershaw, D. Moore, G., Odell, J., et al., *The HTK Book (for HTK Version 3.2.1)*. Entropic Ltd, 2002.