INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2003/M9801
July 2003, Trondheim, Norway

Title: Multi-view 3D-Face Descriptor: proposal for CE

Source: Samsung Advanced Institute of Technology

Author: Won-Sook LEE and KyungAh SOHN

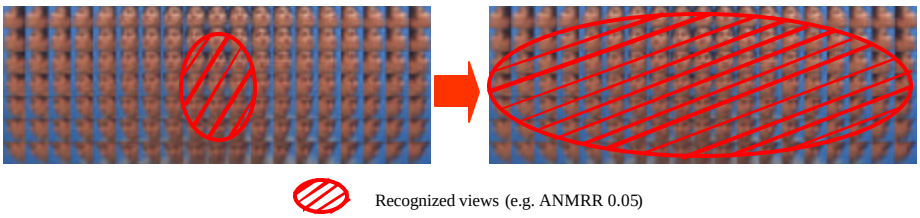
Status: Proposal

1. Introduction

In this contribution, we present a new face descriptor, which aims to contain multi-view 3D-information of a person to help for search, retrieval and browsing of images, videos and 3D-facial model databases.
As described in previous proposal [1], the current model for *Advanced Face Descriptor* [2][3] shows quite satisfactory results to recognize identity of people with given images, and especially frontal view has very high success rate. However as we analyze it, we concludes there is a limitation of the current *Advanced Face Descriptor* definition, which allows only one view to build its descriptor so that it causes high error rate for other views. The source of the problem comes from the fact a face is not a 2D-object, but a 3D-object. To retrieve a 3D-object, we need to give 3D-information to descriptor. In this document, first we do summary of the concept of *Multi-view 3D Face Descriptor*, and then primary experimental results.

2. Multi-view 3D-Face Descriptor

The new descriptor we propose is called as *Multi-view 3D Face Descriptor*, which is supposed to have 3D-information of a face by describing the face as a mosaic of many one-views. This *Multi-view 3D Face Descriptor* aims to raise the success rate to retrieve a face in any view between horizontal rotation [$-90^o \ldots 90^o$] and vertical rotation [$-30^o \ldots 30^o$] ~ [$-60^o \ldots 60^o$] as shown in Figure 1.



(a) *Advanced Face Descriptor* (b) *Multi-view 3D Face Descriptor*

Figure 1: The aim of the *Multi-view 3D Face Descriptor* compared to the current descriptor

To extend the current descriptor to multi-view version, there are many probems to overcome as follows:

1 *DB collection for training/test*
2 Multi-view face detector: Important, but non-normative

3 View estimator: Important, but non-normative
4 Face alignment: Important, but non-normative
5 *Feature extraction*
Modified version of frontal view optimized feature extraction – MPEG-7 *Advanced Face Descriptor* – is used, but as our experiment shows the current model is not strong for the profile views.
6 *Descriptor optimization*
The most naïve idea to create multi-view descriptor is the simple integration of N one-view descriptor. If we register every 10^o apart, we have to register 19 x 7 views if we want to cover horizontal rotation [$-90^o \ldots 90^o$] and vertical rotation [$-30^o \ldots 30^o$]. A very naive descriptor has size 133 x one-view descriptor size. Then the descriptor becomes too big.

As we learn from the current face, a registered view is used to retrieve nearby view with high retrieval rate (e.g. ANMRR) and we extend the concept of quasi-frontal to quasi-view to treat not only frontal view but also any view. Some useful terms are defined as follows:

- ✓ *View-Mosaic*: Mosaic of views 10^o apart covering horizontal rotation [$-X^o \ldots X^o$] and vertical rotation [$-Y^o \ldots Y^o$]. Here we use $X = 90$ and $Y = 30$. It can be visualized as the combination of facial images as shown in Figure 1.

- ✓ *Quasi-view*: The definition of this term is an extension of quasi-frontal, from frontal view to general view. Quasi-view *V'* of a given (registered) view *V*, for instance with ANMRR less than or equal to 0.05, means faces on *V'* are retrieved with *V*, for instance with ANMRR less than or equal to 0.05.

- ✓ *Quasi-view size:* Number of quasi-view with a given view and a given retrieval rate, e.g. with retrieval error less than or equal to 0.05.

3. Localization of faces with angle variation

The localization specification is defined as follows:

- ✓ Size of images: 56x56
- ✓ Positions of two eyes in the front view are on (0.3, 0.32) and (0.7, 0.32) when width and height are considered as 1.0.
- ✓ Left eye position of the positive horizontal rotation keeps (0.3, 0.32) while right eye position of the negative rotation does (0.7, 0.32).
- ✓ Vertical rotation has the same eye positions as the ones on zero vertical rotation images.

To help the understanding, see **Figure 2**.

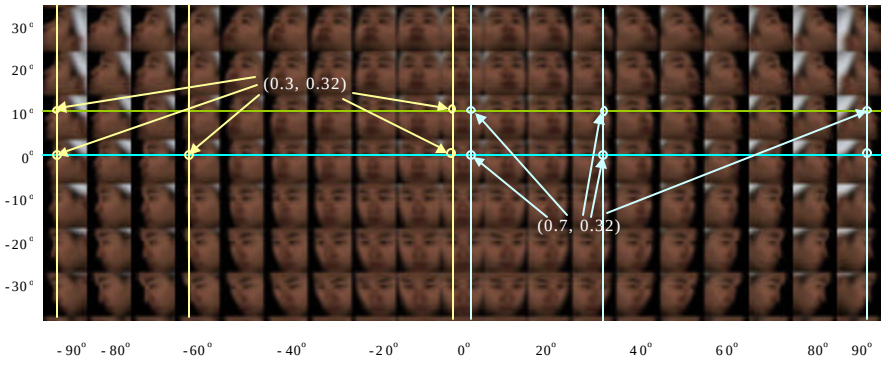


Figure 2: eye-positions on view-mosaic of the center face of rendered images of 108 3D-facial mesh models

4. Experiment with rendered images of 3D facial models

The experiment is done with 50/50 ratios, which means half of images are used for training and the other half for test.

4.1. Feature extraction

We use modified version of the current best algorithm [2], Subregion-based LDA on Fourier space as shown in Figure 3.

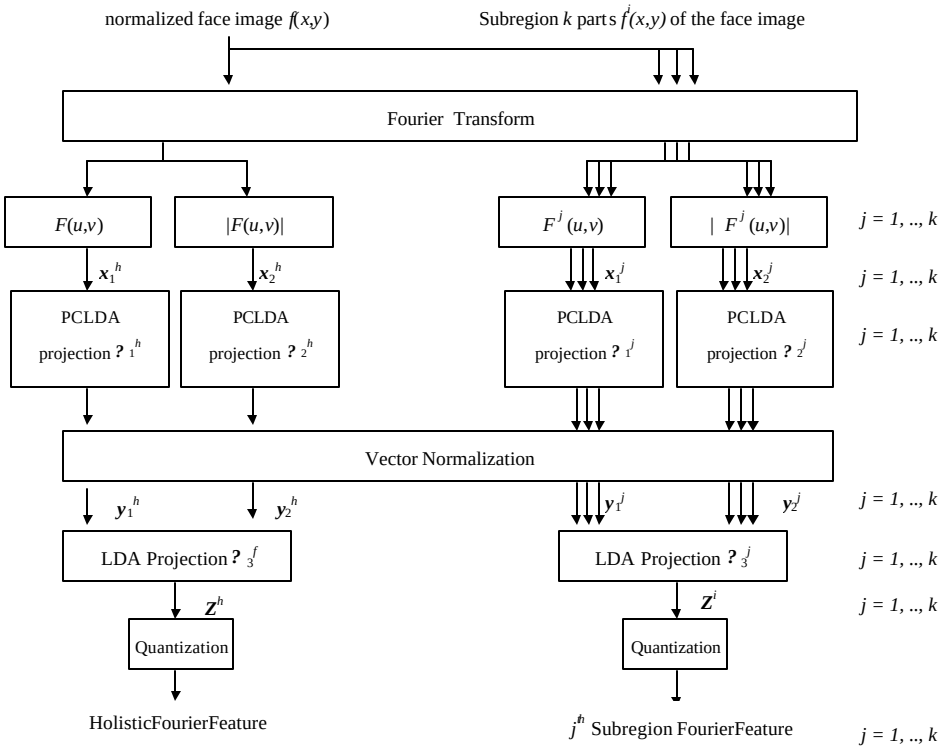
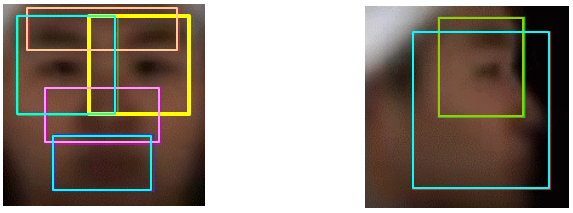


Figure 3: Feature extraction used for *Multi-view 3D Face Descriptor*

Here we vary number and positions of subregions depending on a given view. For example, for the profile view, if we use the same definition of subregion used in the front view, the background may seriously affect for recognition rate. So we define different subregion as shown in Figure 4.



Five subregion definition on View ($0^o, 0^o$) Two subregion definition on View ($80^o, 0^o$)

Figure 4: Subregion definition depending on views superimposed on the center face of our database

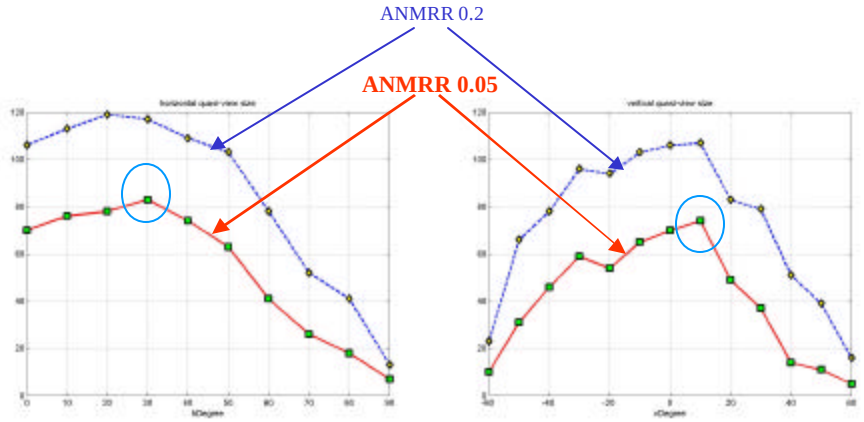
4.2. Quasi-view with 3D model rendered images

Graham and Allinson [4] have calculated the distance between faces over pose to predict the pose dependency of a recognition system. The faces are further apart they will be easier to recognize using distance measures in the eigenspace. Using the average Euclidean distance between the people in the database over the pose angles sampled, they predict that faces should be easiest to recognize around the

30^o range and consequently, the best pose samples to use for an analysis should be concentrated around this range. Additionally they expect that faces are easier to recognize at the frontal view (0^o, 0^o) than the profile (90^o, 0^o). Note that they have checked only horizontal rotation of human heads.

Quasi-view size also depends on the view. Here we consider Quasi-view with ANMRR less than or equal to 0.05. Here **Figure 5** shows how the quasi-view size varies with horizontal and vertical rotations of a head. They are obtained using 108 3D-facial mesh models with 50/50 ratio for training and test. To make fair comparison between different views, we extracted 24 holistic features (without using subregion features) for each view. For horizontal rotation, 9 training views are used for each view from (0^o, 0^o) to (70^o, 0^o), 8 training views for the view (80^o, 0^o) and 7 training views for the view (90^o, 0^o). For vertical rotation, 9 training views are used for each view from (0^o, -40^o) to (0^o, 40^o), 8 training views for the views (0^o, -50^o) and (0^o, 50^o) and 7 training views for the views (0^o, -60^o) and (0^o, 60^o). To help to understand which training views are used for a registered view, see Figure 6.

Figure 5 (a) shows very similar pattern with the graph showing the average distance between faces over view described in Graham and Allinson's paper [4]. The views (20^o, 0^o) ~ (30^o, 0^o) have both the biggest quasi-view size and the biggest Euclidean distance between the people in eigenspace among views (0^o, 0^o), (10^o, 0^o), … , and (90^o, 0^o). **Figure 5** (b) shows the views (0^o, 0^o) ~ (0^o, 10^o) have the biggest quasi-view size among views (0, -60^o), (0^o, -50^o), … , and (0^o, 60^o). The views of heading downward have bigger quasi-view size than ones of heading upward and it makes us to guess it is easier to recognize people when they look downward more than they look upward.



(a) horizontal rotation (b) vertical rotation

Figure 5: Quasi-view sizes with horizontal and vertical rotations with ANMRR 0.05 and ANMRR 0.2

4.3. Descriptor optimization

There are many factors to make the descriptor optimization.

1. Data analysis should be done to make better optimization.
   More frequently appeared poses must be registered or given a large descriptor and seldom appeared poses can be given with small descriptor and neglected. For instance, in an ATM environment, looking down and frontal views are frequent and in Door Access environment, looking down, frontal, side views are frequent. Professional photographer prefers to take photos of a person, not the frontal, but a bit side view and it's the same story in TV camera.
2. We register a few views to produce face descriptor. We select
   - ✓ views with bigger quasi-view size. It's cost-effective to register a view with bigger quasi-view size than one with smaller size.
   - ✓ views which appears a lot in practice.
   - ✓ views where the integration of their quasi-view covers big region of view-mosaic.
   - ✓ views easy to register or easy to obtain.
3. Various feature numbers depending on the view should be used.
   If a view is easy to be obtained for registration, but not so frequently appeared in practice, then we can use a small number of features. More important views get bigger feature numbers.
4. Many views are trained for one registered view to make more robust face descriptor.
   If we can embed more information in the step of training, the registration can be done with smaller information. So we use many views for training for one view registration. For example for the

registered view (30°, 0°), we use 9 views for training such as (10°, 0°), (20°, 0°), (30°, 0°), (40°, 0°), (50°, 0°), (30°, -20°), (30°, -10°), (30°, 10°), (30°, 20°).

4.3.1 Training and Registration views

Training is considered as a processed step to create space basis and matrix transform and we give bigger view information to help wider view region recognition. So instead of using only one view, we use nearby views together. As shown in Figure 6, for one view registration, the training is done with 6 to 9 views around the registered view.

4.3.2 How many training views and features are selected for a registered view

For this experiment, we have given three ways to extract features based on basic feature extraction method described in Section 4.1. Number of subregions and number of features on subregions varies. So for some views, 5 holistic features and 5 features for 5 subregions are extracted and for other views 5 holistic features and 2 features for 5 subregions are extracted. If a view is close to profile, we use 5 holistic features and 5 features for 2. For details for our experiment, see Figure 4 and Figure 6. For one view, one image is selected.

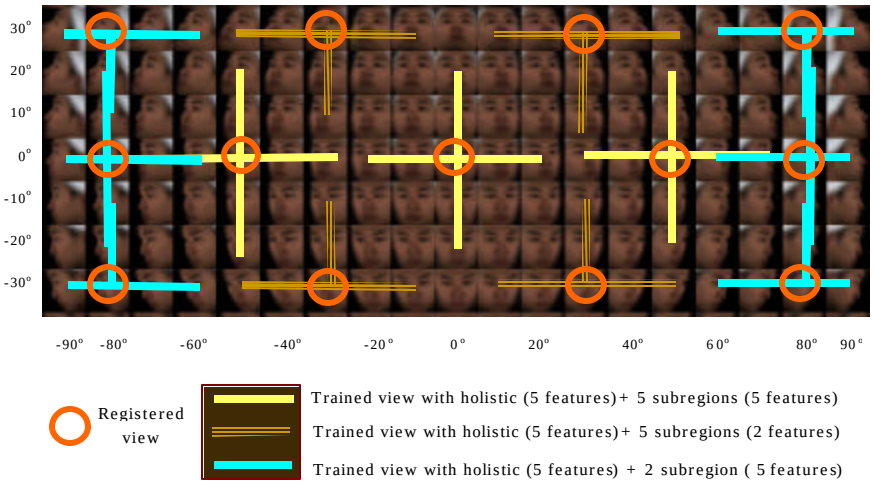


Figure 6: Views used for training and registration.

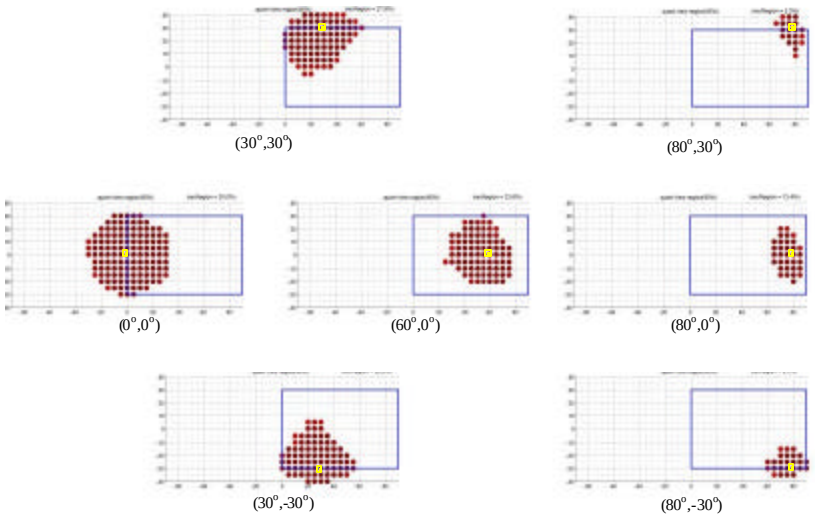


Figure 7: Small rectangle spots are the registered views and small sphere spots indicate corresponding quasi-views with ANMRR less than or equal to 0.05. The large rectangle shows the view region of horizontal rotation [$0^o \ldots 90^o$] and vertical rotation [$-30^o \ldots 30^o$].

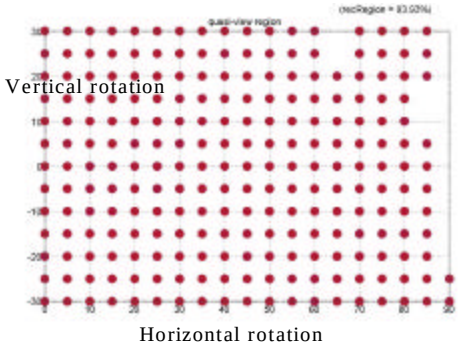


Figure 8: The sphere spots shows the region covered by 7 quasi-views in the view-mosaic of horizontal rotation [0^o… 90^o] and vertical rotation [-30^o… 30^o] with ANMRR less than or equal to 0.05. The covered region is 93.93%.

Through experiments with various combinations of quasi-views, a set of views is selected to create multi-view 3D descriptor. The descriptor has 240 dimensions for rendered images and the detail is shown in Table 1. This descriptor is able to retrieve the rendered images in the test database with ANMRR less than or equal to 0.05 covering 93.93 % views of total region of view-mosaic of horizontal rotation [-90^o… 90^o] and vertical rotation [-30^o… 30^o]. For a reference, it covers 95.36% for ANMRR less than or equal to 0.1, 97.57% for ANMRR less than or equal to 0.15 and 97.98% for ANMRR less than or equal to 0.2.
As a reference, for photograph images, the current *Advanced Face Descriptor* [2][3] has 48 dimensions with ANMRR 0.3013 and 128 dimensions with ANMRR 0.2491 for 50/50 ratio.

View	Extracted features	Feature dimension
(0^o, 0^o)	holistic features + 5 subregion features	1*5 + 5*5 = 30
(60^o, 0^o)	holistic features + 5 subregion features	1*5 + 5*5 = 30
(30^o, 30^o)	holistic features + 5 subregion features	1*5 + 5*2 = 15
(30^o, -30^o)	holistic features + 5 subregion features	1*5 + 5*2 = 15
(80^o, 0^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
(80^o, 30^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
(80^o, -30^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
(-60^o, 0^o)	holistic features + 5 subregion features	1*5 + 5*5 = 30
(-30^o, 30^o)	holistic features + 5 subregion features	1*5 + 5*2 = 15
(-30^o, -30^o)	holistic features + 5 subregion features	1*5 + 5*2 = 15
(-80^o, 0^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
(-80^o, 30^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
(-80^o, -30^o)	holistic features + 2 subregion features	1*5 + 2*5 = 15
Total dimension of features extracted		**Þ 240**

Table 1: Features extracted for *Multi-view 3D Face Descriptor*

5. Comparison between 3D model rendered images and video images

Video streams have been taken with one camera only and semi-automatic localization is processed. For one human seven video streams have been taken. Human subject was rotated horizontally while looking at different height. Figure 9 shows a view-mosaic of the center face of our video face database. Figure 10 shows the quasi-view size compared to one of the rendered images with 15 training human subjects and 16 test subjects. The training data does not have pose information as accurate as the rendered images from the 3D-facial mesh models, especially vertical axis of views. In addition the human subject has been sat for a while and rotated to take video streams, facial, hair and body movement and deformation exist. So the quasi-view size of video images is smaller than the one of rendered images.

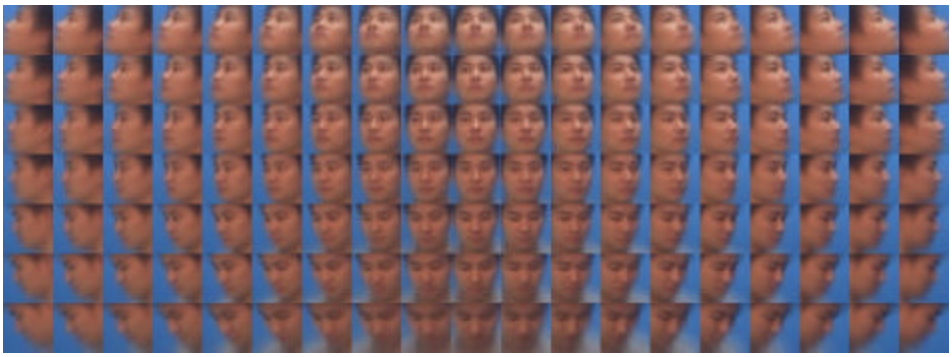


Figure 9: View-mosaic of center face of video images.

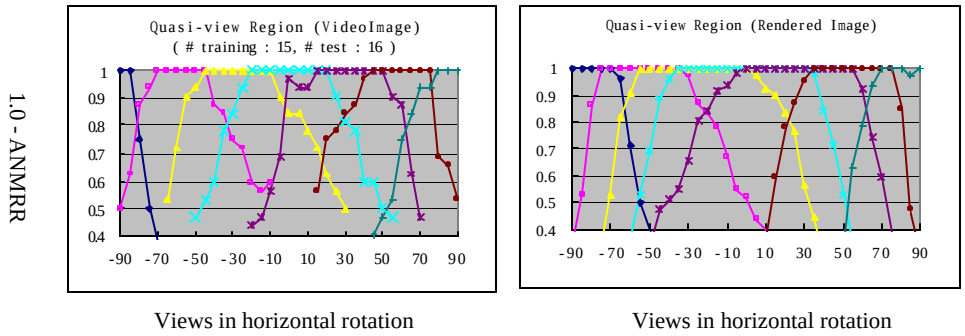


Figure 10: Quasi-view of video images vs. Quasi-view of rendered images of 3D-facial mesh models

6. Conclusion

In this contribution, we experiment of *Multi-view 3D-Face Descriptor* based on the new concept *3D-Face Descriptor* [1]. Among 108 3D-facial mesh models, half of them are used as training and the other half as test. Total 13 views are chosen as registered views. This descriptor with 240 dimension is able to retrieve images of 93.93 % views of total region of view-mosaic of horizontal rotation [$-90^o \ldots 90^o$] and vertical rotation [$-30^o \ldots 30^o$] with accuracy ANMRR less than or equal to 0.05.

The issues in *Multi-view 3D-Face Descriptor* are different from current *Advanced Face Descriptor* in their goal and usefulness. The aim of the new descriptor is to find an optimization to build any-view-information by giving more information and by optimizing them.

There are still many problems to solve and experiment to do such as (i) more video stream image database build-up and experiments for descriptor optimization (ii) the profile view feature extraction (iii) missing view interpolation in the registration step. The interpolation methods with registered views must be considered for the practical use.

Finally, we propose to set up and start a core experiment on the *Multi-view 3D-Face Descriptor*.

7. References

[1] W.-S. Lee S. C. Kee, "3D-Face Descriptor: proposal for CE," ISO/IEC JTC1/SC29/WG11 M9422, March 2003, Pattaya, Thailand, March 2003

[2] A. Yamada and L. Cieplinski, "MPEG-7 Visual part of eXperimentation Model Version 17.1", ISO/IEC JTC1/SC29/WG11 M9502, Pattaya, Thailand, March 2003

[3] T. Kamei, A. Yamada, H. Kim, W. Hwang, T.-K. Kim, S. C. Kee "CE report on Advanced Face Recognition Descriptor", ISO/IEC JTC1/SC29/WG11 M9178, Awaji, JP, December 2002

[4] D. B. Graham and N. M. Allinson, "Characterising Virtual Eigensignatures for General Purpose Face Recognition" in *Face Recognition: From Theory to Applications* (H. Wechsler, PJ Phillips, V. Bruce, FF Soulie and TS Huang, eds.), Berlin: Springer-Verlag, pp. 446-456, 1998