Face Likelihood Functions for Visual Tracking in Intelligent Spaces

Frank Sanabria-Macías, Enrique Marañón-Reyes and Pedro Soto-Vega
Neuroscience, Signals and Images Processing Center
Universidad de Oriente
Santiago de Cuba, Cuba
Email: fsanm77@fie.uo.edu.cu, enriquem@fie.uo.edu.cu

Marta Marrón-Romera, Javier Macias-Guarasa and Daniel Pizarro-Perez
Department of Electronics, Universidad de Alcalá
Alcalá de Henares, Spain
Email: marta@depeca.uah.es, macias@depeca.uah.es
pizarro@depeca.uah.es

***Abstract*—The Viola and Jones face detectors and Particle Filters are great algorithms for face detections and target tracking. However Viola outputs a binary result, while Particle Filters work with probabilistic inputs. This is the reason why there are not so many works that combine both algorithms. A probabilistic model or likelihood functions to transform Viola and Jones output to probabilistic data are needed to allow linking both methods. In this work we explore some Viola and Jones based likelihood functions presented in literature, and propose new strategies. We also extend the evaluation of the likelihood functions in position, scale and pose. One of our proposed functions shows better characteristics to be used in intelligent spaces in three dimensional face tracking applications.**

I. Introduction

Intelligent Spaces are environments equipped with a set of sensorial, communication and computer systems, transparent and imperceptible for the user, that are continuously perceiving the surroundings and cooperating among themselves to provide the necessary help to each person in his/her interaction with the space.

In this context in which a natural interaction between the surroundings and the users is required, it is fundamental to have precise information about the existence and position of the aforementioned users. That is why it is necessary to detect, locate and track each user in the scene [1]. In order to solve these tasks, the information captured by the set of sensors is used.

In practice, many of the intelligent spaces found in the literature use a set of video cameras to accomplish these tasks. One of the most usual approaches involve tracking the users' faces, mainly because it is useful for further tasks such as face recognition, gaze estimation and improving sound location, and the quality of the speech signals generated by the users.

Automatic face detection and tracking in images or sequence of images have been approached using different strategies [2]:

- Some are based on facial features and local characteristics, in which they look for certain elements related to the face elements (feature-based approaches)
- Some are based in holistic approaches, working with certain areas of the full image, extracting features that can represent the searched objects (image-based approaches).

In 2001 Paul Viola and Michael Jones [3] presented a face detection algorithm based in a multistage classification algorithm that substantially improves the processing time, while maintaining high classification rates. After their work, many of the proposed detectors resulted in variants or improvements to the original Viola and Jones algorithm.

In addition, in order to locate faces in image sequences (video), the general strategy aims at first detecting them in each image. Then, and, in order to make good use of the data temporal redundancy, it is common to include tracking algorithms, that model the dynamics of the objectives (faces), so that a reduction of detection and location errors can be achieved.

Particle filters are probabilistic tracking algorithms that allow modeling the system behavior through probability density functions. The essence of these methods consists of propagating *a posteriori* probabilities of the state of the system (face location), given a set of input measurements (for instance the information of the detection algorithm [3]). Particle filters in particular use a set of state-space points called *particles* to model *a posterior* probabilities.

The main difficulty to combine an appearance based face detector, like Viola and Jones, with a tracking algorithm like the particle filters, lies in the fact that the former provides a binary result (face/non-face) while the latter requires probabilistic information. For this reason, in the literature it is not common to find combinations of particle filters with Viola and Jones based face detectors, but with other methods like skin color detectors, in spite of the better performance achieved by the Viola and Jones detectors. This fact has motivated some attempts to combine both algorithms, proposing a probabilistic model, or at least a likelihood function, as the link between Viola and Jones and the probabilistic tracker.

In this work we explore some Viola and Jones based likelihood functions presented in literature, and assess new proposals.

In section II, an algorithmic revision is done, describing both the general Viola and Jones algorithm, and multi-pose and 3D detection strategies. In section, III likelihood functions based on Viola and Jones are described, including a review of

978-1-4799-0224-8/13/$31.00 ©2013 IEEE

previous proposals and the descriptions of the new strategies proposed in this paper. Section IV describes the experimental setup, with the results being discussed in section V. Finally, section VI presents the main conclusions and future work.

II. The Viola and Jones Algorithm

The Viola and Jones algorithm [3] is a powerful and popular algorithm for object detection in images, mainly oriented to face detection applications. It is an appearance based method, which uses statistical methods that allow constructing a face/non-face classifier (a two class pattern recognition problem). The strategy used focuses in detecting light and shadow patterns that can be commonly associated to a resemblance of a human face. The Viola and Jones algorithm uses a set of basic Haar like features and a cascade of classifiers with incremental complexity, each of them being trained with Adaboost.

A. Multi-pose detection strategies

One of the main difficulties of the Viola and Jones algorithm is that classifiers must be trained for specific poses (frontal, left and right sides, etc) with little pose response margin. The different poses a face can present in front of a camera in an intelligent space context, generate the need to have robust algorithms for the detection of these different poses. One of the simplest alternatives to solve this problem is using a set of parallel detectors, each one trained to detect a specific range of pose variations [4]. In our study, we use three trained classifiers, one for frontal pose (F), and two for, left (L) and right (R) side poses. For this purpose, publicly available trained classifiers from the OpenCV library were used [5].

B. 3-D location and tracking

Finally, the 3D localization of faces implies projecting the detected regions in each camera (2D), on the three-dimensional space, and finally combining these regions. However, the binary regions detected by Viola reduce the particle filters effectiveness, as commented above. This fact also limits the possibilities of fusing visual information with other sources like audio, in audiovisual tracking tasks. The other motivation for developing these likelihood functions based on the Viola and Jones scheme is to combine detections from each camera in a probabilistic map, suitable for the 3D particle filter tracking algorithm, thus bringing better precision to the localization task.

III. Viola and Jones based Likelihood Functions

The idea behind this work is being able of generating particles in a 3D space, and finally weighting these particles with 3D probabilistic models. In this model, face likelihood functions will evaluate the projection of particles generated by each camera. In other words, they will be responsible for linking the 2D image, with the 3D face position estimations.

Figure 1 shows a graphical representation of this idea. Particles in 3D represent faces hypothesis located in $X_i = \{x_i, y_i, z_i\}$. This specific location could be some face structure, such as the mouth. Assuming an average human face height these locations can be projected to each camera plane, as a two dimensional bounding box, including position (u, v) and

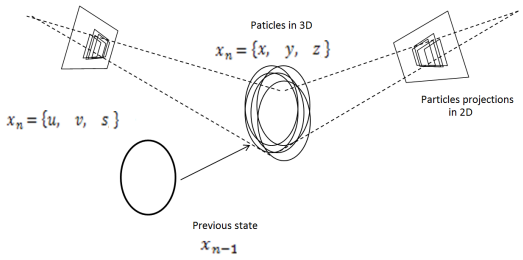

Fig. 1. 3D Localization scheme with likelihood function and motivation

scale (s), fused in a single vector (u, v, s). The likelihood functions generated with the information by each camera can be combined in order to efficiently weight the relevance of each particle. We believe that this strategy can achieve higher performance than those based on skin color detection.

A. Revision of Previous Proposals

In general, the previous attempts to develop likelihood functions based on Viola and Jones, combine the internal information from each classifier in order to build the joint likelihood information. One of the first proposals found in the literature was presented by Li Peihua in 2004 [6]. In that work, the authors propose an empirical expression based on the number of stage classifiers that label the input image as actually being a face, as shown in equation (1):

$$\Omega_{Li1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1-\frac{\lambda}{N_s}}{2\sigma^2}} \tag{1}$$

where λ is the number of stages labeling the input as being a face, N_s is the total number of stages in the classifier and σ is the corresponding standard deviation.

According to this expression, as more stages generate a face label, the probability of this image to be a face will be higher.

In a later work from the same authors [7], a set of five new functions were proposed. The functions are grouped in three classes, named Probability Outer Stages (POS), Probability Inter Stages (PIS) and hybrid stages (PIS-POS), described below. In all the expressions presented next, $f_t^{(k)}$ refers to the response from the t-th weak classifier of the k-th stage. This response will be $\alpha_t^{(k)}$ if this weak classifier labels the input as being a face, and $\beta_t^{(k)}$ otherwise. N_s is the total number of stages, and T_k is the total number of weak classifiers in the k-th stage.

- POS Models: This first group only uses the external information from the stage classifiers. Basically, it uses the number of stages that has passed trough (λ), like in their previous work. The likelihood function is shown in equation (2).

$$\Omega_O(x) = \frac{\sum_{k=1}^{\lambda} \frac{k\phi_k}{N_s}}{\sum_{k=1}^{N_s} \frac{k}{N_s}} \tag{2}$$

where $\phi_k = \begin{cases} 1 & \text{if stage } k \text{ classifies } x \text{ as face} \\ 0 & \text{if stage } k \text{ classifies } x \text{ as not face} \end{cases}$

- PIS-Models: The second group exploits the Adaboost internal structure of the stage classifiers, considering some ratio between how many weak classifiers labeled the input as being a face and the total number of weak classifiers in the stages. The likelihood function are shown in equations (3) and (4).

$$\Omega_{IC}(x) = \frac{1}{N_s} \sum_{k=1}^{N_s} \frac{\sum_{t=1}^{T_k} |\alpha_t^{(k)}|.\delta_t^{(k)}(x)}{\sum_{t=1}^{T_k} |f_t^{(k)}(x)|} \quad (3)$$

$$\Omega_{IE}(x) = \frac{\sum_{k=1}^{N_s} \sum_{t=1}^{T_k} |\alpha_t^{(k)}|.\delta_t^{(k)}(x)}{\sum_{k=1}^{N_s} \sum_{t=1}^{T_k} |f_t^{(k)}(x)|} \quad (4)$$

where $\delta_t^{(k)}(x) = \begin{cases} 1 & \text{if} \quad f_t^{(k)}(x) = \alpha_t^{(k)} \\ 0 & \text{if} \quad f_t^{(k)}(x) = \beta_t^{(k)} \end{cases}$

- PIS-POS Models: The third group tries to combine the strategies of the previous models. The likelihood function are shown in equations (5) and (6).

$$\Omega_{SIO} = \frac{\sum_{k=1}^{\lambda} \frac{k.\phi_k}{N_s}}{\sum_{k=1}^{N_s} \frac{k}{N_s}} \quad (5)$$

$$\Omega_{AIO} = \frac{\sum_{k=N_s-\lambda+1}^{\lambda} \frac{(N_s-k+1).\phi_k}{N_s}}{\sum_{k=1}^{N_s} \frac{k}{N_s}} \quad (6)$$

where $\phi_k = \frac{\sum_{k=1}^{T_k} |\alpha_t^{(k)}|.\delta_t^{(k)}(x)}{\sum_{k=1}^{T_k} |f_t^{(k)}(x)|}$.

Another group of proposed likelihood functions use the expression of conditional probability of detected face given output, within an Adaboost classifier, proposed by Friedman in 1998 [8], and described in equation (7).

$$p(face|H(x)) = \frac{e^{H(x)}}{e^{H(x)} + e^{-H(x)}} \quad (7)$$

where $H(x)$ is the classifier output without applying the corresponding threshold.

Related to this group, are the proposals presented by Boccignone in 2009 and 2010 [9], [10]. They exploit the Bayes' theorem to develop two expressions, one that determines face probability when a given image window passes all stages in the classifier, and another in case it doesn't, as shown in equation (8).

$$\begin{gathered} p(face|F_{N_s}^+, \ldots, F_1^+) = \frac{d^{N_s} p(l=1)}{d^{N_s} p(l=1) + f^{N_s} p(l=-1)} \\ p(face|F_\lambda^-, F_{\lambda-1}^+, \ldots, F_1^+) = \\ = \frac{(1-d)d^{\lambda-1} p(l=1)}{(1-d)d^{\lambda-1} p(l=1) + (1-f)f^{(\lambda-1)} p(l=-1)} \end{gathered} \quad (8)$$

where f is the non face rejection rate, d the face acceptance rate, $p(l=1)$ is the *a priori* probability of finding a face in the classifier input, and $p(l=-1)$ is the *a priori* probability of finding a non-face in the classifier input. The probabilities $p(l=1)$ and $p(l=-1)$ were assumed as being constant and were estimated from the training set. F_k^+ means the windows was classified as face in stage k, while F_k^- means the state k rejects the window.

Other likelihood functions were presented by Li Yuan in 2006 and 2007 [11] [12]. The first combines face and head tracking using head models based on borders and color. As a face detector, a multi-pose detection tree was used instead of a cascade classifier like the Viola and Jones method. In this detection tree, each node is a classifier trained with Adaboost. Like in the cascade approach, a window is classified as face if starting from the root node, it reaches some leaf node. The likelihood function used in this case is shown in equation (9).

$$p(face|x) = \frac{r_v p(H_v(x)|face)}{r_v p(H_v(x)|face) + p(H_v(x)|\overline{face})} \quad (9)$$

where r_v is the *a priori* probability ratio to find a face/non-face window at the classifier input in the leaf node v of the tree.

r_v parameter, can be seen as a fraction of $p(l=1)/p(l=-1)$ in Li Peihua's work, and, like in the previous work, it is obtained from training data. The difference in this case is that this constant is modified from one level to the next in the tree. This modification considers that the deeper we are in the tree, the less probable it is to find a non-face window

Li's work in 2007 [12], combines face detection with the Lukas-Kanade feature based algorithm. As a face detector, they propose using a multi-pose detection tree. The likelihood function used in this case is shown in equation (10).

$$p(face|x) \propto \frac{1}{1 + r_v e^{-H_v(x)}} \quad (10)$$

where r_v is the ratio described above and $H_v(x)$ is the classifier output without applying the corresponding threshold.

It's important to note that in most of these works ([9], [10], [11], [12]) the authors do not evaluate the actual performance of the likelihood function being used as a face detection proposal, but only the final results of the tracking system.

This specific performance information is only presented (graphically) in Wang's work [7]. For each function, the evaluation is carried out over one image with three faces. In their work, the main conclusion was that the hybrid likelihood functions (PIS-POS) showed better performance. However, the assessment was carried out on a too small number of faces, using a fixed scale. So, the statistical significance of the results is questionable and there is no information on the expected behavior when the analysis window varies in size (and this is a fundamental issue in 3-D location and tracking tasks because it can be used to estimate the distance from the camera to the face. Finally, their evaluation didn't consider multi-pose contexts.

Finally, there are other related proposals by Liu in 2003 [13], Verma in 2003 [14] and Li in 2008 [15]. Liu and Li's use only one stage classifier, while Verma uses a Shneiderman and Kanade classifier, instead of Viola and Jones.

B. Proposal of New Likelihood Functions

Analyzing the difficulties found in the previously proposed functions in the literature, and its mathematical expressions, a new set of likelihood functions is proposed in this paper. These new functions are empiric models, mainly inspired by the expressions in the models by Wang. The new functions have been called Ω_{exp1}, Ω_{exp2}, ..., Ω_{exp6}, and their expressions are shown below.

$$\Omega_{exp1} = \frac{\lambda}{N_s} \tag{11}$$

$$\Omega_{exp2} = 1 - \frac{1}{1 + e^{\left(\frac{\sum_{k=1}^{N_s}(H_k(x) - Th_k)}{\sum_{k=1}^{N_s}|H_k(x) - Th_k|}\right)}} \tag{12}$$

$$\Omega_{exp3} = 1 - \frac{1}{1 + e^{\left(\frac{\sum_{k=1}^{\lambda}(H_k(x) - Th_k)}{\sum_{k=\lambda}^{N_s} k}\right)}} \tag{13}$$

$$\Omega_{exp4} = \frac{\lambda}{N_s} \frac{\sum_{k=1}^{\lambda}(H_k(x) - Th_k)}{\sum_{k=\lambda}^{N_s} k} \tag{14}$$

$$\Omega_{exp5} = \frac{\lambda}{N_s} \frac{\prod_{k=1}^{\lambda}(H_k(x) - Th_k)}{\prod_{k=\lambda}^{N_s} k} \tag{15}$$

$$\Omega_{exp6} = \frac{\lambda}{N_s} \sum_{k=1}^{\lambda} \frac{(H_k(x) - Th_k)}{Th_k} \tag{16}$$

where λ is the number of stages classifying a window as face. N_s is the total number of stage. $H_k(x)$ is the stage k classifier output before applied threshold and Th_k the stage k classifier threshold.

Ω_{exp1} expression is similar to Ω_{Li1} as evaluate the fraction of stage that classify a window as face, but without exponential term. The other expressions are based on the difference between weak classifiers sum and stage threshold $(H_k(x) - Th_k)$. We belief that this term is closer to face likelihood that a weighted fraction of weak classifiers outputs, like Wang proposed. Different combinations of this term form expressions from Ω_{exp2} to Ω_{exp6}.

IV. Experimental Setup

With the objective of analyzing the performance of the likelihood functions in the context of a multi-pose 3-D face location task, we implemented our proposals, plus the ones proposed by Wang ($\Omega_{IC}(x), \Omega_{IE}(x), \Omega_O(x), \Omega_{SIO}(x)$, $\Omega_{AIO}(x)$), the probabilistic models used by Boccignone (Ω_B), and Li (Ω_L). In the Li model case, each Viola trained cascade was treated as a single branch tree to apply the concept.

For the evaluation, a set of images from the *Face Pointing 04 Database* [16] [17] were used. The *Face Pointing 04 Database* contains face images from fifteen people, at different poses in pitch rotation (nine positions from $-60°$ (face down) to $60°$ (face up)), and 13 positions in yaw (azimuth) rotation (from $-90°$ (face left) to $90°$ (face right)). The experiments described in this work only use the images with pitch at $0°$, analyzing the variation in azimuth rotation.

A. Likelihood Expected Behavior

In our experiments we used three classifiers, to evaluate the existence of faces in frontal, left and right positions $[L, F, R]$. In this context, where there is more than one cascade classifier, it is necessary to apply the likelihood functions to each of the cascades, obtaining in this case a likelihood vector $[f_L(x), f_F(x), f_R(x)]$. To obtain a multi-pose accurate model, it is required that these functions satisfy certain properties:

- Likelihood functions most present a maximum value point in position when window match a face (this point is reference in each template as mouth position), also when windows size matches face size (scale). Likelihood value must decrease as windows size and position move away from this point. Thus particle's weighting should be right. This bell-like behavior has not to be gaussian since particle filter has no such restriction.
- In addition it's needed not to loose face on intermediate poses, this means that faces between profile (-90 ,+90) and frontal(0) must have high likelihood in at least one template, left(L), frontal(F) or right(R).
- Finally, it's desirable to be able to estimate face pose. To that propose it's necessary likelihood functions should have discriminant power in azimuth angle for each cascade. This is, higher likelihood value for faces in pose near training pose (detection angles region) and lower values in other template training poses (rejection angles region).

B. Experiment I: Evaluating Performance with Position and Scale Changes

For each image, a sliding window was moved over image, in each (u,v) position Viola and Jones detector was applied, and likelihood functions were computed. This procedure was carried out for each cascade (Left, Frontal and Right) obtaining three response maps $[f_L(u,v), f_F(u,v), f_R(u,v)]$. Each cascade $[L, F, R]$ was evaluated against their corresponding poses, $-90°$ for Left, $0°$ for frontal and $+90°$ for right cascade. To analyze likelihood function behavior in scale, the previous procedure was applied with different sliding windows sizes (scales). For each scale maximum likelihood response was considered as response to that scale. Sliding window sizes varied from smaller to bigger than face size in image. At the end of this experiment the functions with bad performance were not used for the second experiment described next.

C. Experiment II: Evaluating Performance with Pose Changes (azimuth)

To evaluate the performance of the likelihood functions against pose (only azimuth), response maps were calculated for faces with different azimuth angles at elevation of zero degree. In each face pose image, maximum likelihood response was considered as likelihood response at that pose for each cascade (template).

V. Results and Discussion

A. Performance against Position Changes

In figure 2 we show an example of the results for functions Ω_{exp1} (a), Ω_{AIO} (b), Ω_{exp4} (c) and Ω_{exp5} (d) with a frontal cascade and varying position shifts.

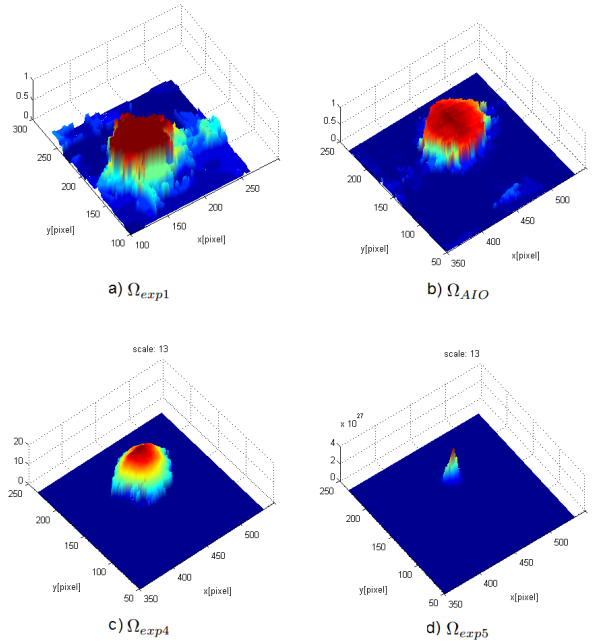

Fig. 2. Results of the evaluation of the likelihood functions with horizontal shifts.

As a result of this evaluation we observed that the functions Ω_{exp1}, Ω_{exp2}, Ω_{IE}, Ω_{IC} and Ω_O showed a saturation behavior in the face region (see figure 2.a) in plane (u, v), without a well defined maximum. Such characteristic is not desirable, for our purposes, because it weighs particles outside the target with high values, making particles centroid shift from the right location, introducing errors in localization task. Besides, this behavior is similar to a binary detector like classic Viola and Jones. This approach has been proved by these authors [18]. Preliminaries test with this bell shaped functions and particles filter have presented lower errors.

Other functions present a bell-like response (see figure 2.b,c,d), with a maximum value in the face area, decreasing its value as the evaluated position gets away from the face area, both in horizontal and vertical directions. In general, the behavior of this likelihood functions, using frontal and profile cascades regarding windows shifts, have similar response.

On the other hand, the function Ω_{AIO} implies evaluating all stages in the classifier for each window, thus significantly increasing its computational cost. Considering that Ω_{AIO} and Ω_{SIO} have similar results, the Ω_{AIO} function was discarded.

B. Performance against Scale Changes

As a result of this evaluation we observed that the likelihood functions Ω_{exp1}, Ω_{exp2}, Ω_{exp3} and Ω_O behavior with scale changes, present the same saturated pattern discussed in the previous section. The Ω_{IC} and Ω_{IE} functions present their maximums at the correct scale, but they also showed high values at lower scales. In general, the other functions presented a better performance in this test: a maximum value near the correct face size, and decreasing value as the evaluated scale moved away from this scale. However, the side face cascades presented a lower decreasing slope as the scale increased, in comparison to the frontal cascade.

C. Performance against Pose Changes

Figures 3 and 4 show likelihood functions response vs pose angle, in the three cascade Left (left column), Frontal (central column) and Right (right column)). The horizontal axis represents the azimuth angle from $-90°$ to $+90°$ while vertical axis is the maximum likelihood value at that pose.

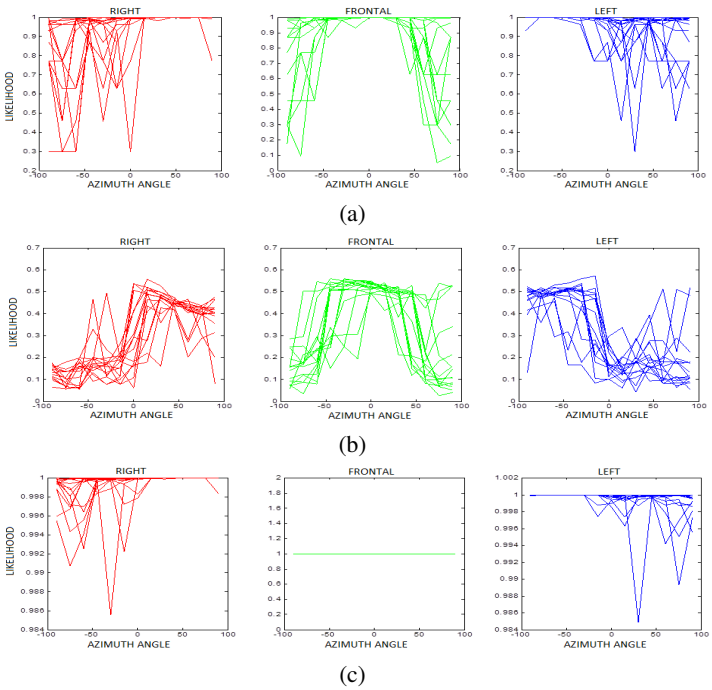

Fig. 3. Likelihood vs Pose evaluations. (a)Ω_B, (b)Ω_{SIO} (c)Ω_L

These figures show that all functions fulfill the second criterium, where face with intermediate pose has high likelihood value in at least one cascade. However figure 3 shows that the behavior of functions Ω_B and Ω_L differs from the desired behavior presented in the third criterium. Although the angles close to the training pose angle (detection angles) exhibit high likelihood values, their variance is high outside this region (corresponding to the rejection angles), with values in the same range as those obtained for the detection angles. Hence these functions are not adequate for discriminating poses, so they were discarded.

The Ω_{SIO} function exhibited a better performance, and we can see the difference between the average values in the detection and rejection angles regions. However the variances in both regions remained high, and showed a saturation like pattern, which is not desired in our task.

The performance of the likelihood functions Ω_{exp3}, Ω_{exp4} and Ω_{exp6}, proposed in this paper (shown in Figure 4) are closer to the desired behavior. These functions present a clear

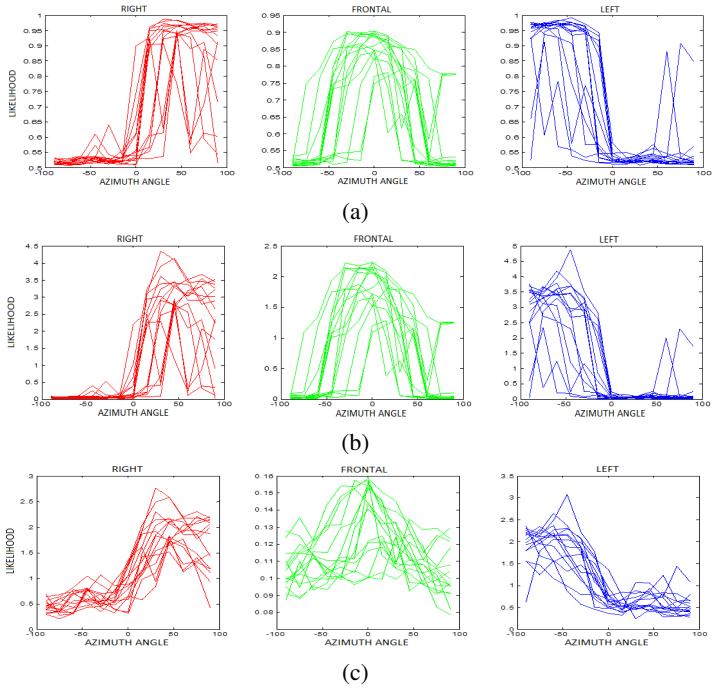

Fig. 4. Likelihood vs Pose evaluations. (a)Ω_{exp3} and (b)Ω_{exp4} (c) Ω_{exp6}

differentiation between the values in the detection and rejection regions for each cascade classifier.

The rejection regions show, in all cases, a low variance and despite the lack of smooth transitions between the detection's and rejection regions, they are smoother than Ω_{SIO}. The model Ω_{exp6} is the one with a smoother transition between regions, but its variance in the rejection zone is higher than for Ω_{exp3} and Ω_{exp4}. It is important to note that the range of likelihood values is not the same for frontal and side cascades. This characteristic is caused by the lack of normalization in the proposed expressions and the fact that the number of stages is not the same in all cases.

VI. Conclusions and Future Work

We have proposed and evaluated several alternatives for likelihood functions to be used in multi-pose face detection tasks. The comparison with previously proposed functions show that our proposal achieved better performance.

We also show that the proposed function Ω_{exp4} has good characteristics for the task. It's behavior results in a smooth gaussian like response against position and pose shifts in the image. It is also able to correctly cover the azimuth pose variations range by combining three classifiers trained for frontal and side poses.

The proposal also shows a discriminative response to pose variations, and this characteristic is interesting when attempting to estimate the face pose in 3D, starting from the results of the three classifiers. Thus, our proposal is then able to transform the Viola and Jones binary classification output into a probabilistic measure, which can be efficiently integrated in a particle tracker.

Future work aims to develop a model that combines the probabilistic output of the face detection likelihood function into a full 3-D face detection and tracking system. Additional effort will also be devoted to design a scheme to combine the results returned by the model in 2D and weight the 3D particles filter from all views. Additionally it is intended that the model and the scheme are able to also estimate the face pose.

References

[1] A. M. party Interaction (AMI) project, "State of the art overview: Localization and tracking of multiple intelocutors with multiple sensors," Augmented Multi-party Interaction (AMI) project, Tech. Rep., 2007.

[2] J. Meynet, T. Arsan, J. Mota, and J. Thiran, "Fast multiview face tracking with pose estimation," Technical report TR-ITS. 2007.01. Ecole Polytechnique Federale de Lausanne (EPFL), Signal Processing Institute, Tech. Rep., 2007.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE CVPR 2001*. Citeseer, 2001.

[4] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 446–453.

[5] "Opencv web page," http://opencv.willowgarage.com/wiki/ [last access April 2013].

[6] P. Li and H. Wang, "Probabilistic face tracking using boosted multi-view detector," in *PCM (2)*, 2004, pp. 577–584.

[7] H. Wang, P. Li, and T. Zhang, "Novel likelihood estimation technique based on boosting detector," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3. IEEE, 2005, pp. III–477.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, p. 2000, 1998.

[9] G. Boccignone, P. Campadelli, A. Ferrari, and G. Lipori, "Real-time probabilistic tracking of faces in video," in *Proceedings of the 15th International Conference on Image Analysis and Processing*, ser. ICIAP '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 672–681.

[10] ——, "Boosted tracking in video," *Signal Processing Letters, IEEE*, vol. 17, no. 2, pp. 129–132, 2010.

[11] Y. Li, H. Ai, C. Huang, and S. Lao, "Robust head tracking based on a multi-state particle filter," 2006.

[12] Y. Li, C. Huang, and H. Ai, "Tsinghua face detection and tracking for clear 2007 evaluation," in *The CLEAR 2007 Evaluation*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds., vol. 4625. Springer, 2007, pp. 138–147.

[13] F. Liu, X. Lin, S. Li, and Y. Shi, "Multi-modal face tracking using bayesian network," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, oct. 2003, pp. 135 – 142.

[14] R. Verma, C. Schmid, and K. Mikolajczyk, "Face detection and tracking in a video by propagating detection probabilities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1215–1228, 2003.

[15] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1728–1740, 2008.

[16] "Pointing'04 icpr workshop, cambridge, united kingdom - 22 august 2004," http://www-prima.inrialpes.fr/Pointing04/data-face.html [last access April 2013].

[17] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004, pp. 1–9.

[18] F. Sanabria-Macías, J. Macías-Guarasa, M. Marrón-Romera, D. Pizarro, and E. Marañón-Reyes, "Seguimiento audiovisual de locutor usando un filtro de partículas extendido con proceso de clasificación," in *SAAEI11. Seminario Anual de Automática, Electrónica Industrial e Instrumentación*, 2011.