Proposition d'une stratégie de fusion de données à trois niveaux pour la vérification d'identité

Patrick Verlinde[*] *Dominique Genoud*[†] *Guillaume Gravier* *Gérard Chollet*

Ecole Nationale Supérieure des Télécommunications / Département Signal
46, rue Barrault
75634 Paris Cedex 13, FRANCE

ABSTRACT

The aim of this paper is to propose a strategy that uses data fusion at three different levels to gradually improve the performance of an identity verification system. In a first step temporal data fusion can be used to combine multiple instances of a single (mono-modal) expert to reduce its measurement variance. If the system performance is too low, one can try to improve it by fusing in a second step the results of multiple experts working on the *same* biometric modality. This supposes that their respective classification errors are (at least partially) uncorrelated. Finally, if the performance is still not good enough, one will be forced to pass on to the third step in which performance can be increased when using multiple experts working on different (biometric) modalities. These experts have to be chosen in such a way that the effect of adding the extra modalities is to increase the separation between the distributions of the different populations that have to be classified by the system. This kind of strategy allows a user to gradually tune the performance of an identity verification system to its needs while controlling the increase of investment costs[1].

1. INTRODUCTION

L'objectif d'un système de v´erification d'identit´e est de d´ecider si une personne qui proclame une identit´e est bien la personne qu'elle pr´etend être. Chaque m´ethode impl´ement´ee, appel´ee ici expert, effectue un test d'hypothèse binaire. Ces tests g´enèrent deux types d'erreurs: un faux rejet (FR: rejeter un accès client) ou une fausse acceptation (FA: accepter un imposteur). Les performances d'un système de v´erification d'identit´e sont en g´en´eral caract´eris´ees par les taux d'erreurs globaux mesur´es pendant les tests: le taux de faux rejets [False Rejection Rate: FRR = (nombre de FR) / (nombre d'accès client)] et le taux de fausses acceptations [False Acceptance Rate: FAR = (nombre de FA) / (nombre d'accès imposteur)]. Une mesure unique peut être obtenue en combinant ces deux erreurs pour obtenir le taux d'erreurs global [Total Error Rate: TER = (nombre de FA + nombre de FR) / (nombre total d'accès)] ou par son compl´ement, le taux de r´eussite global [Total Success Rate: TSR = 1 - TER] [BC97]. Une v´erification d'identit´e parfaite, FAR = 0 et FRR = 0, peut difficilement être obtenue en pratique. Cependant, une ´etude d´etaill´ee des tests d'hypothèse binaire [Tre68] montre que si l'on fait tendre FRR vers z´ero en modifiant le seuil de d´ecision, on augmente alors le taux de fausses acceptations, et inversement.

Les erreurs sont caus´ees par l'imperfection des diff´erents experts. On a donc des erreurs de mesure. Une solution, d´etaill´ee dans la section 2, pour rem´edier à ce problème consiste à combiner les r´esultats obtenus par un seul expert sur des instances successives sur plusieurs tests. On parlera alors de fusion temporelle. De cette manière on peut effectivement diminuer la variance des r´esultats, ce qui permet d'augmenter les performances du système de base sans avoir à modifier l'approche choisie. Si ces performances ne satisfont pas les besoins de l'utilisateur, il peut passer à l'am´elioration suivante.

Le fait de disposer d'experts parfaits et de connaître les distributions statistiques des r´esultats obtenus par chaque m´ethode de v´erification pour les deux classes (clients, imposteurs), permet d'utiliser la règle de d´ecision bay´esienne pour obtenir un fonctionnement du système au taux d'erreur minimal [DH73]. L'erreur de classification ne sera ´egale à z´ero que dans le cas où les deux distributions ne se recouvrent pas du tout. Si cette condition n'est pas satisfaite, ce qui est le cas en g´en´eral, on ne peut faire baisser les deux types d'erreurs (FAR, FRR) *simultanément* qu'en augmentant la taille de l'´echantillon statistique, c'est-à-dire en utilisant plus de donn´ees d'entraînement, permettant ainsi de diminuer les variances des estimateurs des paramètres r´eels de la distribution. Très souvent on ne dispose malheureusement que d'un nombre limit´e de donn´ees d'entraînement, ce qui exclut cette possibilit´e th´eorique. Une autre possibilit´e pour faire diminuer ces deux taux d'erreurs simultan´ement consiste alors à fusionner les r´esultats de diff´erents experts bas´es sur la même modalit´e. Ce type de fusion permet d'augmenter les performances si les erreurs commises par les experts ne sont pas totalement corr´el´ees. L'am´elioration des performances sera même d'autant plus grande que les r´esultats (et les erreurs) obtenus par les diff´erents experts seront plus d´ecorrel´es, l'apport d'information augmentant avec la d´ecorrelation. Dans la section 3, on montrera le cas d'un système mono-modal multi-expert bas´e sur la modalit´e biom´etrique vo-

[*] Ecole Royale Militaire / D´epartement ELTE, Bruxelles, Belgique, e-mail: patrick.verlinde@tele.rma.ac.be

[†] Institut Dalle Molle d'Intelligence Artifi cielle Perceptive, Martigny, Suisse, e-mail: genoud@idiap.ch

[1] Some of the results given in this paper were obtained in the framework of the CAVE-PICASSO and M2VTS European projects.

cale. Dans le cas sp´ecifique d'une modalit´e vocale, on a ajout´e la possibilit´e de diff´erer la d´ecision après fusion en cas de doute. Cette technique permet de diminuer les erreurs au prix d'une augmentation du nombre de non-d´ecisions et d'un traitement suppl´ementaire. Ce m´ecanisme peut ´egalement être utilis´e pour les autres systèmes de v´erification d'identit´e, mais cette augmentation des performances induit des coûts suppl´ementaires. Si les performances du système après fusion mono-modale ne satisfont toujours pas l'utilisateur, il est possible de proc´eder à l'´etape suivante.

Dans un troisième temps on peut esp´erer augmenter encore les performances en trouvant des caract´eristiques qui am´eliorent la s´eparabilit´e des hypothèses test´ees en augmentant la dimension de l'espace. Il est possible d'utiliser à cette fin des modalit´es biom´etriques suppl´ementaires. La discrimination entre les deux distributions sera d'autant plus facile que la correlation entre les diff´erentes modalit´es sera plus faible. On ´etudiera ce cas dans la section 4 pour des modalit´es biom´etriques vocales et visuelles.

2. FUSION MONO-MODALE MONO-MÉTHODE

2.1. Méthodes utilisées

Le premier pas dans la strat´egie propos´ee, est d'utiliser une int´egration temporelle afin d'augmenter les performances d'un système de v´erification d'identit´e en diminuant la variance des mesures. Un exemple de cette technique est pr´esent´e dans [KMJS97]. Les auteurs de cet article partent, d'une part, du principe que les d´ecisions d'accepter (ω_1) ou de rejeter (ω_2) la personne, sont bas´ees sur la probabilit´e de classes *a posteriori* $P(\omega_j|x_i), j = 1, 2$ et d'autre part, du fait que l'on dispose de plusieurs instances x_i de donn´ees biom´etriques. Dans ce cas, on peut se baser sur la fusion de l'information additionnelle. On suppose que les mesures x_i ont ´et´e acquises dans les mêmes conditions et qu'elles peuvent être consid´er´ees comme des mesures multiples qui ne diffèrent entre elles que par leur composante de bruit. Aussi les probabilit´es de classes *a posteriori* peuvent être consid´er´ees comme des estimations bruit´ees de la valeur nominale de cette probabilit´e, *i.e.*:

$$P(\omega_j|x_i) = P(\omega_j|x) + \epsilon(\omega_j|x_i).$$

Ainsi on peut obtenir de meilleures estimations $\hat{P}(\omega_j|x)$ de cette probabilit´e *a posteriori* en combinant les estimations bruit´ees de manière lin´eaire ou en utilisant des statistiques de rang. Les diff´erentes strat´egies de fusion sont compar´ees entre elles en analysant leur probabilit´e d'erreur de manière th´eorique. Les m´ethodes de fusion ´etudi´ees sont les suivantes:

Moyenne arithmétique

$$\hat{P}(\omega_j|x) = \frac{1}{R}\sum_{i=1}^{R} P(\omega_j|x_i).$$

Règle maximale

$$\hat{P}(\omega_j|x) = \max_{i=1}^{R} P(\omega_j|x_i).$$

Règle minimale

$$\hat{P}(\omega_j|x) = \min_{i=1}^{R} P(\omega_j|x_i).$$

Règle médiane

$$\hat{P}(\omega_j|x) = \mathrm{med}_{i=1}^{R} P(\omega_j|x_i).$$

2.2. Résultats

On d´emontre que la r´eduction de la variance est la plus spectaculaire pour la moyenne arithm´etique près de la frontière de d´ecision. Les r´eductions de la variance pour les statistiques de rang d´ependent, entre autres, du nombre d'instances R et de l'ordre du rang et, bien qu'elles soient moins efficaces que dans le cas de la moyenne arithm´etique, elles sont plus robustes par rapport aux valeurs extrêmes des r´esultats de mesure. La table 1 montre une r´eduction des taux d'erreurs jusqu'à 40%. Le gain en performance est monotone au d´ebut d'une s´equence, mais tend à saturer après int´egration des premières observations. Ces r´esultats ont ´et´e obtenus en utilisant un expert de v´erification visuelle, bas´e sur une correlation robuste optimis´ee. Pour ces experiences, l'expert utilise uniquement la modalit´e image de la base de donn´ees multimodale M2VTS [PV97] qui contient 37 personnes.

Table 1: EER en fonction du nombre d'instances

Expert	1	2	3	4	5	6
Moyenne	6.9	4.6	4.5	3.7	4.0	4.0
M´edian	6.9	4.8	4.2	4.3	4.2	4.0

3. FUSION MONO-MODALE MULTI-MÉTHODE

3.1. Méthodes utilisées

La combinaison de plusieurs experts sur une même modalit´e biom´etrique est très utilis´ee. On peut citer par exemple [Ben92, XKS92].

A titre d'exemple plus ´elabor´e, les experiences d´ecrites dans [GBGC96], bas´ees sur plusieurs m´ethodes de v´erification du locuteur en mode d´ependant du texte, sont decrites. les trois experts utilis´es acceptent en entr´ee un ensemble de vecteurs de 12 coefficients cepstraux LPC extraits d'un signal de parole. Ces experts sont:

1. *Dynamic Time Warping (DTW)* L'algorithme de DTW consiste à effectuer une comparaison dynamique entre une matrice de r´ef´erence et une matrice de test. Le r´esultat est une mesure de distance entre le test et la r´ef´erence.
2. *Statistique du second ordre (SSO)* Une matrice de covariance $X = \frac{1}{M}\sum_{t=1}^{t=M} X_t X_t^T$ de la s´equence de parole de r´ef´erence est cr´e´ee. Une matrice de covariance Y de la s´equence de test est ´egalement cr´e´ee. Une mesure de sph´ericit´e sym´etrique $\mu(X, Y)$ est ensuite effectu´ee.

3. *Modèles de Markov cachés (HMM)* Deux sortes de modèles HMM sont cr´e´es pour chaque mot à tester: (1) **Un modèle du monde**, cr´e´e à partir d'une base de donn´ees (Polyphone [CCCL95]). Les modèles de monde sont identiques pour tous les clients. (2) **Un modèle du client**, qui utilise comme paramètres initiaux le modèle du monde, et dont on r´eestime les paramètres pour chaque locuteur avec les donn´ees de celui-ci. La mesure utilis´ee consiste à calculer le log du rapport de vraisemblance locuteur/monde pour un segment de parole donn´e.

Détermination des seuils La d´ecision d'acceptation ou de rejet d'un locuteur s'effectue par comparaison du score de chaque methode à un *seuil a priori*. Le seuil est d´etermin´e par la m´ethode de Furui. Ce seuil est robuste lorsque très peu de donn´ees d'entraînement sont à disposition [Fur81].

Combinaison des décisions Afin d'am´eliorer la d´ecision globale de notre système, nous avons combin´e les d´ecisions de chaque m´ethode (DTW, SSO, HMM). La combinaison utilis´ee ici est un système majoritaire pond´er´e avec seuil de doute. La d´ecision par rapport au seuil *a priori* est pond´er´ee (entre 0 et 1) en fonction de la distance seuil/score, ce qui revient à ´etablir une confiance dans cette d´ecision. Le système compare la moyenne des confiances des m´ethodes qui ont obtenu la majorit´e à un seuil (seuil de doute). Si cette moyenne est sup´erieure au seuil, on prend la d´ecision de la majorit´e, sinon il y a doute.

3.2. Résultats

La base de donn´ees Polycode [GC95] utilis´ee pour ces exp´eriences est constitu´ee de 25 locuteurs. La table 2 montre que la fusion des d´ecisions des trois experts est meilleure que le meilleur des experts. Pris dans un cadre plus formel, les trois experts utilis´es ne sont pas fondamentalement diff´erents, la DTW et les SSO ´etant des cas "d´eg´en´er´es" de la mod´elisation HMM. Cependant la mod´elisation des locuteurs effectu´ee par chaque expert n'´etant pas parfaite, ce sont parfois les algorithmes n´ecessitant moins de paramètres à estimer qui s'avèrent les plus robustes. Parmi les facteurs qui entachent l'estimation correcte d'un modèle, le manque de donn´ees d'entraînement et la variabilit´e intra-locuteur sont certainement les plus importantes.

Table 2: Performance des experts et de la fusion

Expert	FRR% (200 tests)	FAR% (1800 tests)
DTW	23.5	7.67
SOSM	14.0	5.28
HMM L/R	5.53	2.72
Fusion	2.0	2.72

Malgr´e l'am´elioration des performances due à la fusion, les erreurs r´esiduelles peuvent, dans certaines applications, s'av´erer inacceptables, et il est possible, dans ce cas, d'´eviter de prendre de mauvaises d´ecisions en rejetant les cas douteux. La table 3 nous indique le comportement des performances du système lors du changement de seuil de doute. Ce seuil permet de r´egler ainsi les taux de faux rejets et fausses acceptations selon le niveau de s´ecurit´e requis par l'application. Les cas douteux peuvent être r´esolus en demandant des informations suppl´ementaires au locuteur.

Table 3: Performance en changeant le seuil de doute

Seuil de doute	FRR% (200 tests)	FAR% (1800 tests)	Doute% (2000 tests)
0.2	2.0	2.72	0.0
0.5	2.0	2.33	0.83
0.7	2.0	1.11	10.11
0.8	1.0	0.89	20.2

4. FUSION MULTI-MODALE MULTI-MÉTHODE

4.1. Méthodes utilisées

Le troisième pas dans la strat´egie de d´eveloppement propos´ee est la fusion de multiples experts utilisant plusieurs modalit´es (biom´etriques). Cette approche a, entre autre, ´et´e utilis´ee dans les r´ef´erences suivantes [BF95, DB97, JLGW97].

Pour montrer les possibilit´es de la fusion multi-modale, nous avons test´e diff´erents modules de fusion et nous avons compar´e les r´esultats à ceux obtenus avec les experts individuels, sur la base de donn´ees multi-modale M2VTS [PV97] avec cinq experts:

UCL Pro expert visuel bas´e sur les profils (Universit´e Catholique de Louvain);

UCL Grey expert visuel bas´e sur les profils et les niveaux de gris (Universit´e Catholique de Louvain);

RMA Pro expert visuel bas´e sur les profils (Ecole Royale Militaire);

UOS Cor expert visuel frontal bas´e sur une m´ethode de corr´elation robuste (University of Surrey);

RMA Ar expert vocal bas´e sur une m´ethode auto-r´egressive (Ecole Royale Militaire).

On a test´e les modules de fusion suivants: m´ethodes par vote (ET, OU, Majoritaire) [Das94], classificateurs lin´eaire et quadratique [DH73] et le perceptron multi-couches (MLP) [Bis95].

4.2. Résultats

Les r´esultats exp´erimentaux se trouvent dans la table 4. Ces r´esultats, exprim´es en taux d'´egales erreurs, ont ´et´e obtenus en utilisant les seuils *a posteriori* et repr´esentent donc les meilleures performances possibles.

Après ce premier test, on a compar´e la meilleure m´ethode avec un MLP (caract´eris´e par 25 neurones logsig dans la

Table 4: Taux d´egales d'erreurs avant et après fusion

Expert	EER%
UCL Pro	08.0
UCL Grey	07.0
RMA Pro	17.0
UOS Cor	12.0
RMA Ar	02.0
M´ethode par vote ET	12.5
M´ethode par vote OU	06.5
M´ethode par vote MAJ	05.0
Classificateur lin´eaire	02.5
Classificateur quadratique	01.0

couche cach´ee et un neurone logsig dans la couche de sortie), sur des donn´ees de test en utilisant les seuils calcul´es *a posteriori* (selon le critère EER). Les r´esultats se trouvent dans la table 5. On constate que la fusion de donn´ees multi-modale permet effectivement d'augmenter les performances au delà des performances du meilleur expert.

Table 5: Comparaison des erreurs de v´erification

M´ethode	FRR%	FAR%
Classificateur quadratique	0.0	4.1
MLP	0.0	0.4

5. CONCLUSIONS

La fusion de donn´ees peut être utilis´ee au sein d'un système de v´erification d'identit´e à diff´erents niveaux. Le premier niveau est la fusion temporelle des r´esultats obtenus par un seul expert, et donc sur une seule modalit´e, afin de r´eduire la variance de mesure. Un deuxième niveau est atteint en fusionnant les r´esultats obtenus par diff´erents experts sur la même modalit´e, ceci afin de minimiser les erreurs de classification en jouant sur la d´ecorrelation des erreurs des diff´erents experts. Le troisième niveau d'application est de diminuer encore les erreurs de classification en essayant d'accroitre la distance entre les distributions des deux populations en augmentant la dimension de l'espace par l'utilisation de plusieurs modalit´es (biom´etriques) les plus d´ecorrel´ees possible. Nos travaux actuels sont ax´es vers une combinaison de ces trois utilisations de la fusion de donn´ees dans une seule strat´egie s´equentielle, qui a le grand avantage d'augmenter les performances d'un système de v´erification d'identit´e de manière graduelle.

BIBLIOGRAPHIE

[BC97] F. Bimbot and G. Chollet. Assessment of speaker verification systems. In *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, 1997.

[Ben92] Y. Bennani. Text-independent talker identification system combining connectionist and conventional models. In S. Y. Kung et al., editor, *Neural Networks for Signal Processing, Vol.2*. IEEE Service Center Press, 1992.

[BF95] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford UK, 1995.

[CCCL95] G. Chollet, J.L. Cochard, A. Constantinescu, and Ph. Langlais. Swiss french polyphone and polyvar: telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, 1995.

[Das94] B. V. Dasarathy. *Decision Fusion*. IEEE Computer Society Press, 1994.

[DBB+97] B. Duc, E. Bigün, J. Bigün, G. Maître, and S. Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):835–843, 1997.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[Fur81] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on ASSP*, ASSP-29(2):254–272, 1981.

[GBGC96] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet. Combining methods to improve speaker verification decision. In *Proceedings of The Fourth International Conference on Spoken Language Processing*, 1996.

[GC95] Dominique Genoud and G´erard Chollet. Polycode a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

[JLGW97] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[KMJS97] J. Kittler, G. Matas, K. Jonsson, and M. U. Ramos S´anchez. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[PV97] S. Pigeon and L. Vandendorpe. The M2VTS multi-modal face database. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[Tre68] H. L. Van Trees. *Detection, Estimation and Modulation Theory*, volume 1. John Wiley & Sons, New York, 1968.

[XKS92] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, May/June 1992.