Εξαγωγή Μεταδεδομένων από Βίντεο Κινήσεων Νοηματικής Γλώσσας μέσω Δικτύου Bayes

Ι. Ασκάρογλου[α], Στ. Τζικόπουλος[α], Δ. Κοσμόπουλος[β], Σ. Θεοδωρίδης[α]

[α] Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Πανεπιστημιούπολη, 157 84, Ιλίσια, Αθήνα, Ελλάδα
[β] Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών "ΔΗΜΟΚΡΙΤΟΣ", Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, 153 10, Αγία Παρασκευή, Ελλάδα

Περίληψη

Σε αυτή την εργασία επιχειρείται η εξαγωγή μεταδεδομένων από βίντεο νοηματικής γλώσσας μέσω επεξεργασίας βασισμένης σε όραση (vision - based). Τα μεταδεδομένα αυτά αφορούν την αναγνώριση της ταυτότητας των περιοχών δέρματος που ανιχνεύονται, δηλαδή κεφάλι, χέρια καθώς και επικαλύψεις αυτών, οι οποίες είναι απαραίτητες για υψηλότερου επιπέδου επεξεργασία.

Αρχικά, εξάγονται τα χαμηλού επιπέδου χαρακτηριστικά των περιοχών δέρματος. Με χρήση των κατάλληλων ετικετών – ταυτοτήτων των περιοχών αυτών εκπαιδεύεται ένα Bayesian δίκτυο για να γεφυρώσει το χάσμα μεταξύ της υψηλού επιπέδου γνώσης για την δομή του ανθρωπίνου σώματος και τα χαμηλού επιπέδου χαρακτηριστικά με μία πιθανοτική μέθοδο, επιτρέποντας την επίλυση του δύσκολου προβλήματος των επικαλύψεων.

Η προσέγγιση αυτή εφαρμόζεται σε βίντεο νοηματικής γλώσσας, τα οποία προήλθαν από δική μας αναπαραγωγή κινήσεων της Αμερικανικής Νοηματικής Γλώσσας (American Sign Language - ASL). Μετά από πειράματα στα βίντεο αυτά, το ποσοστό αναγνωρισιμότητας των περιοχών ανήλθε στην τάξη του 90%, εξάγοντας πολύτιμα συμπεράσματα για την μέθοδο, η οποία μπορεί να γενικευθεί σε οποιαδήποτε άλλη περίπτωση, όπου προκύπτει υψηλού επιπέδου πληροφορία από χειρονομία.

Keywords: Χειρονομίες Νοηματικής Γλώσσας, Ιστόγραμμα, Ροπές Zernike, Εκπαίδευση Bayesian Δικτύου, Αναγνώριση Περιοχής Δέρματος

1. Εισαγωγή

Το πρόβλημα της αναγνώρισης νοηματικής γλώσσας είναι αρκετά σύνθετο λόγω της ιδιόμορφης και μη διεθνοποιημένης φύσης της, αφού δεν υπάρχει μία διεθνής νοηματική γλώσσα. Σημαντικό ρόλο στην αποκωδικοποίηση των γλωσσών αυτών παίζουν όχι μόνο οι κινήσεις χεριών και κεφαλιού αλλά και εκφράσεις προσώπου και μετακίνηση όλου του σώματος. Μερικές φορές η κατανόηση κάποιων κινήσεων είναι αδύνατη χωρίς να λάβουμε υπόψη μορφασμούς προσώπου, κοίταγμα ματιών και άλλες κινήσεις του σώματος. Τα συστήματα αναγνώρισης κινήσεων νοηματικής γλώσσας γενικά διακρίνονται σε δύο ειδών: τα βασισμένα α) σε όραση (vision-based)

και β) σε όργανα μέτρησης (instrumentation-based). Εδώ θα ασχοληθούμε μόνο με την πρώτη μέθοδο, η οποία χρησιμοποιεί κάμερες που αποθηκεύουν την κάθε εικόνα (frame) με έναν σταθερό ρυθμό δειγματοληψίας (sampling rate) και ύστερα το σύστημα τις επεξεργάζεται. Τα βασικά προβλήματα που συναντώνται στην πορεία αυτής της διαδικασίας είναι: η ποικιλία της τοποθέτησης στο χώρο του προτύπου μιας κίνησης, η συγχώνευση των κινήσεων λόγω της ταχύτατης διαδοχής τους και η επικάλυψη (occlusion) των μελών του σώματος που συμμετέχουν σε μία κίνηση. Για τους λόγους αυτούς η αυτόματη αναγνώριση κινήσεων (automatic gesture recognition) μέσω οπτικής μόνο επεξεργασίας είναι μεγάλη πρόκληση.

Όσον αφορά τις σχετικές με το θέμα μελέτες που έχουν γίνει στο παρελθόν, οι προσπάθειες ήταν πολλές και με διαφορετικές μεθόδους, αλλά χωρίς να επικρατήσει κάποια από αυτές. Μία από τις πιο πρόσφατες είναι η [Lee et. al., 1999], όπου εκπαιδεύεται ένα Hidden Markov Model. Εργασία με άμεση σχέση με την παρούσα είναι η [Derpanis, 2003] με τη διαφορά ότι προηγείται της αναγνώρισης μια ενδελεχής γλωσσολογική μελέτη της νοηματικής γλώσσας. Μία δημοφιλής μέθοδος είναι αυτή της οπτικής ροής (Optical Flow) ([Cutler et. al., 1998], [Allen et. al., 1993]), όπου η επεξεργασία γίνεται επί της διαφοράς διαδοχικών εικόνων. Μέθοδος που έχει χρησιμοποιηθεί στη [Ahuja, 1998] είναι η τεχνική βασισμένη στην τροχιά (Trajectory-based), ενώ παραλλαγή της παραπάνω μεθόδου ([Bashir et. al., 2005]) χρησιμοποιεί γκαουσιανά μοντέλα ανάμειξης (Gaussian Mixture Models - GMM). Μελέτες που εφαρμόζουν Bayesian δίκτυα (Bayesian Network -BN-) είναι η [Park et. al., 2004] και η [Gong et. al., 2002]. Το βασικό τους μειονέκτημα παρόλα αυτά είναι ότι δεν αντιμετωπίζουν αποτελεσματικά τις επικαλύψεις μελών του σώματος και συγκεκριμένα χεριών μεταξύ τους και χεριών με το κεφάλι. Εξαίρεση αποτελούν αυτές, που αντιμετωπίζουν το πρόβλημα με κατάλληλο σχηματισμό στερεοσκοπικής κάμερας ή με χρήση τρισδιάστατου μοντέλου [Rehg et. Al, 1993]. Γενικά οι μέθοδοι που έχουν εφαρμοστεί μέχρι τώρα χρησιμοποιούν χαμηλού επιπέδου χαρακτηριστικά για να συμπεράνουν σημασιολογία υψηλότερου επιπέδου, παραλείποντας όμως την αξιοποίηση της γνώσης υψηλότερου επιπέδου. Αυτό επιτυγχάνεται με την εφαρμογή Bayesian δικτύων μέσω επιπέδων επεξεργασίας και πιθανοτικής μοντελοποίησης.

2. Κυρίως Κείμενο

2.1. Τεχνική Προσέγγιση

Στην αρχή του αλγορίθμου και μόνο για το πρώτο frame, ανιχνεύεται το πρόσωπο, το οποίο αναπαρίσταται από υποσύνολο συναρτήσεων κυματιδίων χρησιμοποιώντας κατάλληλα τα περιστρεφόμενα χαρακτηριστικά της συνάρτησης κυματιδίου Haar. Ύστερα επιλέγοντας μία πρότυπη περιοχή γύρω από τα μάτια προκύπτουν τρία ιστογράμματα του χρωματικού μοντέλου HSV, τα οποία χρησιμοποιούνται ως μοντέλο δέρματος. Κάθε ιστόγραμμα προσεγγίζεται με γκαουσιανή καμπύλη και γίνεται αναζήτηση περιοχων στην εικόνα με παραπλήσια ιστογράμματα. Παράλληλα

χρησιμοποιούνται εύρωστες στατιστικές μέθοδοι ελαχιστοποίησης σφάλματος ([Kosmopoulos et. al., 2006]).

Σαν τελική επεξεργασία πριν την εξαγωγή των χαρακτηριστικών απορρίπτονται οι περιοχές θορύβου που έχουν ανιχνευθεί λανθασμένα ως δέρμα. Αυτό γίνεται με την επιλογή το πολύ τριών περιοχών (χέρια και κεφάλι) από αυτές που έχουν βρεθεί, τις τρεις μεγαλύτερες από πλευράς εμβαδού. Για κάθε περιοχή υπολογίζεται ένα διάνυσμα χαρακτηριστικών. Δύο από αυτά προκύπτουν από την θέση της περιοχής (απόσταση και γωνία) σε σχέση με το κέντρο βάρους και των τριών περιοχών που έχουν βρεθεί στο συγκεκριμένο frame. Όλα τα υπόλοιπα προκύπτουν από ροπές Zernike (complex Zernike moments) μέχρι και $4^{ης}$ τάξης. Η ροπή Zernike τάξης m ($A_{m,n}$, όπου $|n| \leq m$ και $m - |n|$ = άρτιος αριθμός , με m να ανήκει στους μη μηδενικούς ακεραίους και n σε όλους τους ακεραίους) δίνεται με πολικές συντεταγμένες από τις εξισώσεις 1 έως 3, σύμφωνα με τις [Kosmopoulos et. al., 2006] και [Shutler et. al., 2002].

$$A_{m,n} = \frac{m+1}{\pi} \cdot \int_0^1 \int_{-\pi}^{\pi} f(r,\theta) \cdot [V_{m,n}(r,\theta)^*] dr d\theta \quad \textbf{(1)}$$

$$V_{m,n}(r,\theta) = R_{m,n}(r) e^{jn\theta} \quad \textbf{(2)}$$

$$R_{m,n}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s \frac{(m-s)!}{s!(\frac{m+|n|}{2})!(\frac{m-|n|}{2}-s)!} r^{m-2s} \quad \textbf{(3)}$$

Όσο μεγαλύτερη είναι η τάξη των ροπών, που χρησιμοποιούμε, τόσο πιο ακριβής θα είναι η ανακατασκευή της εικόνας, αλλά και τόσο πολυπλοκότερη θα είναι υπολογιστικά η διαδικασία. Η πληροφορία, που περιέχεται σε υψηλότερης τάξης ροπές, περιλαμβάνει περισσότερες λεπτομέρειες της εικόνας με αποτέλεσμα την αύξηση της μεταβλητότητα μεταξύ διαφορετικών εικόνων ακόμα και του ίδιου ατόμου. Συνεπώς ενώ στην φάση της συλλογής των δεδομένων υπολογίστηκαν οι ροπές μέχρι $7^{ης}$ τάξης, για την εκπαίδευση του δικτύου χρησιμοποιήθηκαν μόνο μέχρι $4^{ης}$. Αφού συλλεχθούν οι τιμές των χαρακτηριστικών για όλα τα βίντεο που είχαμε στην διάθεσή μας, ακολουθεί η κβάντισή τους, ώστε να εκπαιδευτεί το BN.

Τα Bayesian δίκτυα είναι ένα σύνολο από τυχαίες μεταβλητές X={X1, X2,..., XN} και αποτελούνται από μία Δικτυακή Δομή S και ένα σύνολο P συναρτήσεων κατανομής πιθανότητας που κάθε μία αντιστοιχεί σε μία από τις N μεταβλητές. Η δικτυακή δομή είναι ουσιαστικά ένας μη κυκλικός κατευθυνόμενος γράφος, όπου η έλλειψη ακμών σημαίνει υπό συνθήκη ανεξαρτησία των μεταβλητών-κόμβων. Οι σχέσεις των υπό συνθήκη πιθανοτήτων μας δίνουν τη δυνατότητα να υπολογίσουμε

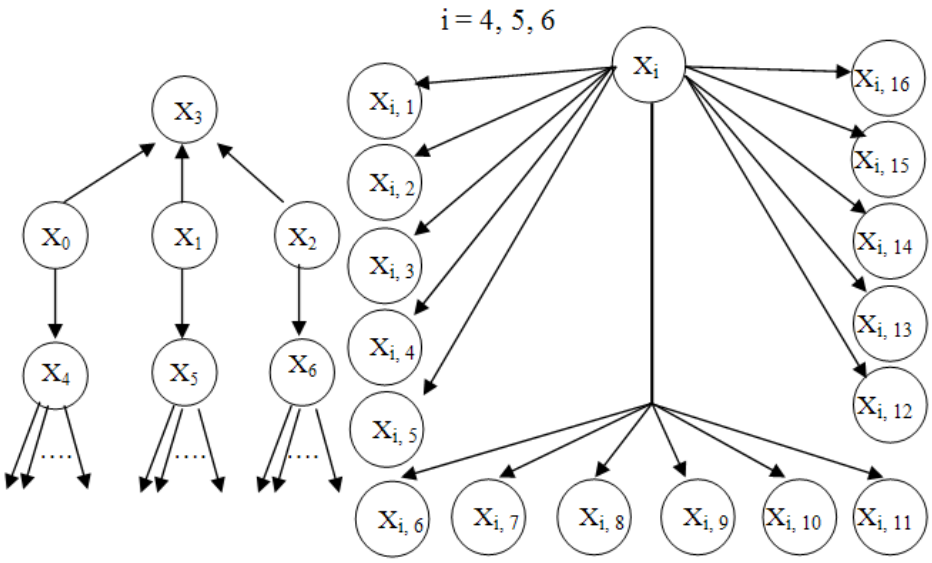
την κατανομή της από κοινού πιθανότητας (Joint Probability Distribution) P(U). Ένα Bayesian δίκτυο μας παρέχει την δυνατότητα να υπολογίσουμε την τιμή v της πιθανότητας μιας μεταβλητής V έχοντας ως δεδομένες τις τιμές e ενός συνόλου μεταβλητών E (Evidence), δηλαδή υπολογίζουμε την πιθανότητα P(V=v|E=e). Αυτό σημαίνει ότι παρακολουθώντας το σύνολο μεταβλητών E και γνωρίζοντας τις τιμές τους e υπολογίζω την πιθανότητα η μεταβλητή V να έχει τιμή v, δεδομένου του e. Για περισσότερες πληροφορίες παραπέμπουμε στις αναφορές [Murphy, 1998], [Heckerman, 1995], [Huang et. al., 1996] και [Pearl, 1986]. Το Bayesian δίκτυο αρχικά εκπαιδεύεται με δεδομένες τιμές χαρακτηριστικών και ετικετών (φάση εκπαίδευσης) και στη συνέχεια δοκιμάζεται η ικανότητα αναγνωρισμότητας περιοχών σε άγνωστα βίντεο (φάση δοκιμής).

Το σημασιολογικό μοντέλο που χρησιμοποιούμε αποσκοπεί στην πιθανοτική μοντελοποίηση της γνώσης της δομής του ανθρώπινου σώματος ώστε να εξάγει μεταδεδομένα. Η μελέτη [Gong et. al., 2002] επεκτείνεται εφαρμόζοντας ένα Bayesian δίκτυο το οποίο μοντελοποιεί ξεχωριστά τις περιπτώσεις επικάλυψης (occlusions) καθώς και οι μεταβλητές που αποτελούν το δίκτυο προκύπτουν από τον υπολογισμό Zernike ροπών, όπως γίνεται και στην παρούσα εργασία.

Δεδομένου συνόλου μεταβλητών x, που αναπαριστώνται σαν κόμβοι στο δίκτυο, αναζητούμε το κατάλληλο διάνυσμα αυτών των μεταβλητών, ώστε να μεγιστοποιηθεί η από κοινού κατανομή πιθανότητας (Joint Probability Distribution) δοσμένων κάποιων παρατηρήσεων e (evidence), οι οποίες προέρχονται από υπολογισμούς της εικόνας και είναι συνυφασμένοι με άλλους κόμβους του δικτύου. Η τιμή του διανύσματος αντιπροσωπεύει την τρέχουσα κατάσταση της χειρονομίας, δηλαδή την θέση του κεφαλιού και των χεριών στην εικόνα καθώς και τις επικαλύψεις. Η δομή του δικτύου που εφαρμόστηκε απεικονίζεται στο σχήμα 1.

Αναλυτικά οι μεταβλητές-κόμβοι του δικτύου και τι αντιπροσωπεύουν παρουσιάζονται παρακάτω:

- X_0, X_1, X_2: Μεταβλητές που αντιστοιχούν στις (το πολύ) τρεις περιοχές ενδιαφέροντος εκφράζοντας την πιθανότητα ότι μία αυτές είναι δεξί ή αριστερό χέρι, κεφάλι, επικάλυψη κάποιου είδους (δεξί ή αριστερό χέρι με κεφάλι, τα δύο χέρια με κεφάλι και τα δύο χέρια μεταξύ τους) ή απλώς θόρυβος. Οι τιμές των μεταβλητών αυτών αποτελούν την «απόφαση» του δικτύου κατά την φάση της δοκιμής.
- X_3: Δυαδική μεταβλητή που εξαρτάται από τις τιμές των X_j, j=0,1,2. Χρησιμοποιείται για να αποκλείονται οι μη δυνατοί συνδυασμοί των μεταβλητών X_j, j=0,1,2 λόγω της δομής του ανθρώπινου σώματος.
- X_4, X_5, X_6: Πρόκειται για βοηθητικές μεταβλητές που «αποδεσμεύουν» τους κόμβους X_0, X_1, X_2, X_3 από τους γράφους του σχήματος 1.β. Χρησιμοποιούνται μόνο στην εκπαίδευση του δικτύου.



Σχήμα 1. Δομή Bayesian δικτύου: α) ανώτερο επίπεδο και β) κατώτερο επίπεδο

Καθώς όπως είπαμε οι κόμβοι X_j, j=4,5,6 αντιστοιχούν σε μία περιοχή δέρματος και έχουν «γονείς» τους X_j, j=0,1,2, οι κόμβοι-παιδιά του σχήματος 1.β λειτουργούν ως δεδομένα (evidence) κατά την εκπαίδευση του δικτύου και οι τιμές τους είναι:

- $X_{i,1}$, $X_{i,2}$: Δηλώνουν το μέτρο και την γωνία αντιστοίχως του διανύσματος AB, όπου A το κέντρο βάρους της τρέχουσας περιοχής i και B το κέντρο βάρους όλων των περιοχών δέρματος.
- $X_{i,16}$: Η ταυτότητα της περιοχής στο προηγούμενο frame. Μοντελοποιεί την συνέχεια της κίνησης.
- $X_{i,3}$, $X_{i,4}$, $X_{i,5}$, $X_{i,7}$, $X_{i,9}$, $X_{i,11}$, $X_{i,12}$, $X_{i,14}$: Αντιστοιχούν στην Ευκλείδια νόρμα των ροπών Zernike μέχρι 4ης τάξης της περιοχής i από το κέντρο βάρους της. Να αναφέρουμε ότι το μέτρο της ροπής Zernike με m=n=1 είναι πάντα μηδέν ($|A_{11}=0|$).
- $X_{i,6}$, $X_{i,8}$, $X_{i,10}$, $X_{i,13}$, $X_{i,15}$: Αντιστοιχούν στις μη μηδενικές γωνίες (φάσεις) των παραπάνω Zernike ροπών.

Στη δομή του Bayesian δικτύου έχουμε 4 επίπεδα. Τα δεδομένα (evidence) του δικτύου προέρχονται από τους κόμβους-παιδιά του $4^{ου}$ επιπέδου, οι οποίοι και μοντελοποιούν την χαμηλού επιπέδου πληροφορία που μας παρέχει κάθε εικόνα-frame. Ομοίως, η μεταβλητή-κόμβος X_3 του $1^{ου}$ επιπέδου μοντελοποιεί την υψηλού επιπέδου γνώση που κατέχουμε για την δομή του ανθρωπίνου σώματος και χρησιμοποιείται στην διαδικασία της δοκιμής, επιτρέποντας συγκεκριμένους συνδυασμούς τιμών των μεταβλητών του κάτω επιπέδου της. Αυτή η δομή υλοποιεί τον συνδυασμό χαμηλού (μέσω του $4^{ου}$ επιπέδου), μεσαίου (μέσω του $3^{ου}$ και $2^{ου}$ επιπέδου) και υψηλού επιπέδου γνώσης (μέσω του $1^{ου}$ επιπέδου), στον οποίο

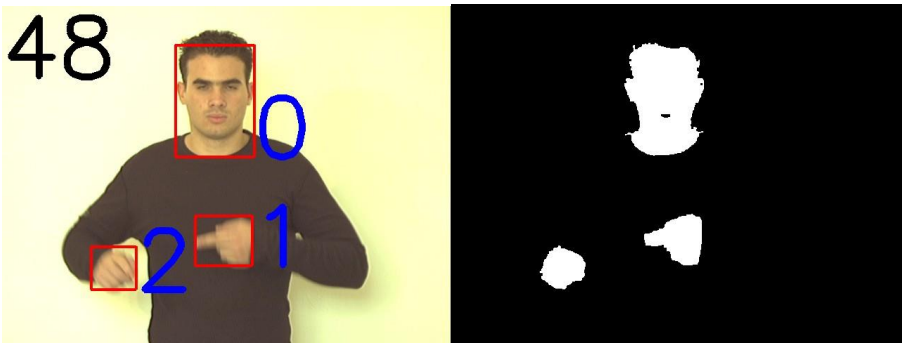
αναφερθήκαμε στην Εισαγωγή επιτρέποντας την εξαγωγή μεταδεδομένων. Με αυτό τον τρόπο προκύπτουν αποτελέσματα με υψηλά ποσοστά αναγνωρισιμότητας.

Μετά την εκπαίδευσή του, πειραματιζόμαστε με άγνωστα βίντεο, ώστε να βγάλουμε συμπεράσματα για την αποτελεσματικότητα του δικτύου (φάση δοκιμής). Το δίκτυο καλείται να εξάγει σημασιολογία μεσαίου επιπέδου, δηλαδή τι περιοχή του σώματος (χέρι, κεφάλι, συγκεκριμένη επικάλυψη αυτών ή μεταξύ τους, ή θόρυβος) είναι οι περιοχές δέρματος που ανιχνεύει. Η υλοποίηση των παραπάνω βημάτων και αλγορίθμων έγινε με χρήση δοσμένου κώδικα, ο οποίος όμως υπέστη σημαντικές μεταβολές, που διευκρινίζονται στο πειραματικό μέρος.

2.2 Πειραματικό Μέρος

Στο πρώτο κομμάτι του πειραματικού μέρους, συγκεντρώσαμε σε μορφή βίντεο όλες τις κινήσεις της αμερικανικής νοηματικής γλώσσας (American Sign Language – ASL), από τον διαδικτυακό τόπο *www.commtechlab.msu.edu*. Έπειτα δημιουργήσαμε μία Βάση Δεδομένων Βίντεο, αναπαράγοντας 20 επιλεγμένες κινήσεις από την ASL από 4 διαφορετικά άτομα-εθελοντές. Οι κινήσεις αυτές επιλέχθηκαν ώστε στο σύνολό τους να έχουν μία ποικιλία ως προς την κινησιολογία τους. Το κάθε άτομο εκτέλεσε κάθε κίνηση 10 φορές. Τα 800 βίντεο που καταγράφτηκαν, σε μορφή αλληλουχίας εικόνων jpeg, χρησιμοποιήθηκαν στο μετέπειτα στάδιο της εκπαίδευσης και δοκιμής του Bayesian δικτύου και είναι διαθέσιμα στην ιστοσελίδα *http://www.iit.demokritos.gr/~dkosmo/downloads/gesture*.

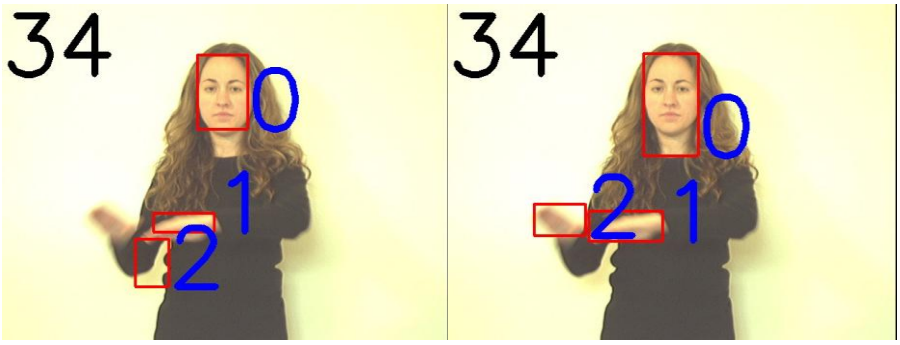
Έχοντας, πλέον, κατασκευάσει τη βάση δεδομένων, περνάμε στη φάση της εξαγωγής των δεδομένων εκπαίδευσης. Σε κάθε frame κάθε βίντεο εκτελείται ο αλγόριθμος που έχουμε ήδη αναφέρει, εντοπίζονται οι περιοχές δέρματος (σχήμα 2), υπολογίζεται το διάνυσμα χαρακτηριστικών κάθε περιοχής και δίνεται με χειροκίνητη μεθοδολογία το είδος της (χέρι, κεφάλι, συγκεκριμένη επικάλυψη αυτών ή μεταξύ τους, θόρυβος).



Σχήμα 2. α) Ανιχνευμένες περιοχές του βίντεο knife4a, β) Η αντίστοιχη μάσκα

Κατά τη διάρκεια του βήματος αυτού αντιμετωπίσαμε τα ακόλουθα προβλήματα: α) φυσικά ενιαίες περιοχές δέρματος ανιχνεύονται «κομμένες», δηλαδή σαν δύο μικρές

αντί μία ενιαία, β) περιοχές δέρματος δεν ανιχνεύονται καθόλου ή ανιχνεύονται σε πολύ μικρότερο εμβαδό και γ) περιοχές της εικόνας που ανιχνεύονται να μην αποτελούν δέρμα, όπως φαίνεται και στο σχήμα 3.α. Τα προβλήματα αυτά προκύπτουν από το γεγονός ότι κατά την προσέγγιση των ιστογραμμάτων με γκαουσιανές καμπύλες, όπως επίσης και κατά την σύγκριση των περιοχών με αυτές τις καμπύλες υπεισέρχονται σφάλματα. Επομένως, τα ιστογράμματα των περιοχών που θα θέλαμε να εντοπίζει το σύστημα δεν εντάσσονται στα όρια που επιτρέπουν οι γκαουσιανές αυτές καμπύλες. Εξετάζοντας όμως το θέμα της προσέγγισης του ιστογράμματος με γκαουσιανή καμπύλη πιο αυστηρά και επεμβαίνοντας περισσότερο σε αυτή τη διαδικασία, παρατηρήσαμε, ότι το πρόβλημα εντοπιζόταν στο κανάλι Value. Ορίζοντας πλέον εμείς εξωτερικά στο σύστημα τις παραμέτρους κάθε γκαουσιανής καμπύλης, ανιχνεύονται οι σωστές περιοχές. Αυτό προϋπέθετε κάποιο πειραματισμό, έδωσε όμως ικανοποιητικά αποτελέσματα, όπως απεικονίζεται και από το σχήμα 3.β.



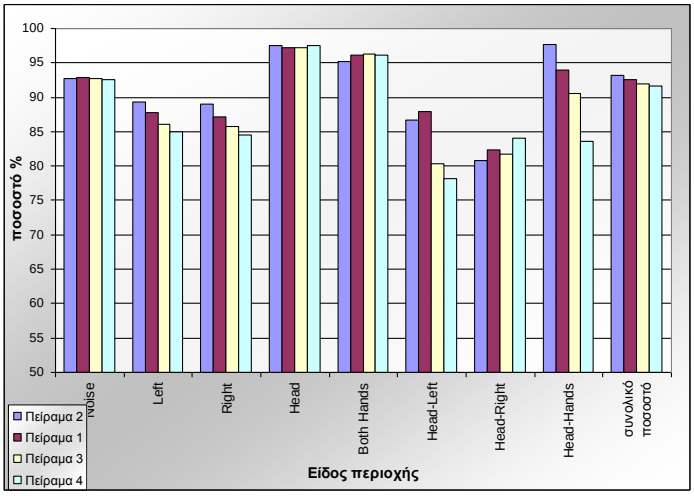
Σχήμα 3. Ανιχνευμένες περιοχές δέρματος: α) Αρχική προσέγγιση, β) Βελτιωμένη προσέγγιση

‘Εχοντας πλέον υπολογίσει όλα τα απαραίτητα δεδομένα για κάθε περιοχή δέρματος, επόμενο βήμα είναι να χρησιμοποιηθεί μέρος των δεδομένων αυτών για την εκπαίδευση του Bayesian δικτύου και το υπόλοιπο κομμάτι για το στάδιο του ελέγχου του Δικτύου. Προκειμένου όμως τα δεδομένα αυτά να χρησιμοποιηθούν ως δεδομένα εκπαίδευσης του δικτύου θα πρέπει να κβαντιστούν. Η κβάντιση αυτή ανά διάνυσμα-χαρακτηριστικό είναι απαραίτητη, διότι στη μελέτη αυτή χρησιμοποιείται δίκτυο με διακριτές τιμές μεταβλητών. Για να κβαντιστούν τα δεδομένα που έχουμε με όσο το δυνατόν μικρότερο σφάλμα, για δεδομένο πλήθος τιμών κβάντισης, χρησιμοποιήθηκε ο αλγόριθμος Lloyd. Τα μέτρα των ροπών Zernike κβαντίζονται στις 10 τιμές ενώ οι γωνίες τους στις 8.

2.3 Πειράματα - Συμπεράσματα

Εφόσον έχουν υπολογιστεί και κβαντιστεί όλα τα απαραίτητα δεδομένα για κάθε περιοχή δέρματος, χρησιμοποιούμε ένα ποσοστό των δεδομένων αυτών για την

εκπαίδευση του Bayesian δικτύου. Το υπόλοιπο ποσοστό δίνεται στο δίκτυο ώς άγνωστες περιοχές, με σκοπό την αναγνώρισή τους από το δίκτυο και την εξαγωγή ποσοστών επιτυχίας. Ξεκινώντας θελήσαμε να ελέγξουμε την εξάρτηση του πλήθους των δεδομένων εκπαίδευσης με το ποσοστό σωστά αναγνωρισμένων περιοχών δέρματος. Στο σχήμα 4 απεικονίζονται τα αποτελέσματα των τεσσάρων πρώτων πειραμάτων. Στα πειράματα αυτά, τα δεδομένα εκπαίδευσης είναι ισομερώς κατανεμημένα ως προς τα πρόσωπα αλλά και τις κινήσεις. Το ποσοστό των δεδομένων εκπαίδευσης ως προς το συνολικό ποσοστό των δεδομένων κυμαίνεται από 80% (Πείραμα 2) έως 10% (Πείραμα 4). Μειώνοντας το ποσοστό των δεδομένων εκπαίδευσης, μειώνεται και το ποσοστό των ορθά αναγνωρισμένων περιοχών δέρματος από 93.21% (Πείραμα 2) σε 91.65% (Πείραμα 4), παραμένοντας όμως πάντα σε υψηλά επίπεδα.

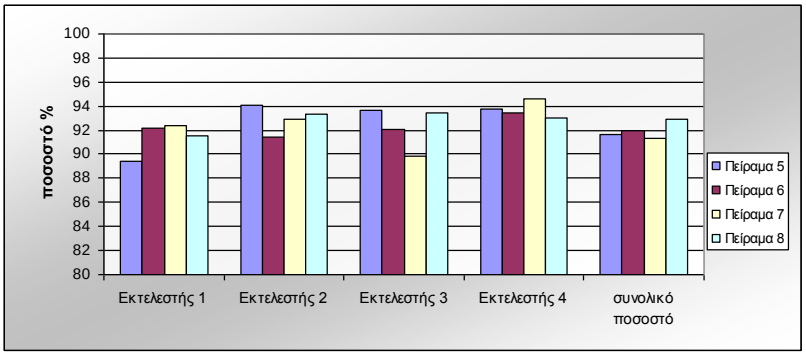


Σχήμα 4: Ποσοστό % σωστά αναγνωρισμένων περιοχών ανά είδος περιοχής των πειραμάτων 1-4

Όσον αφορά τις εσφαλμένα αναγνωρισμένες περιοχές δέρματος, μπορούμε να παρατηρήσουμε τα εξής:

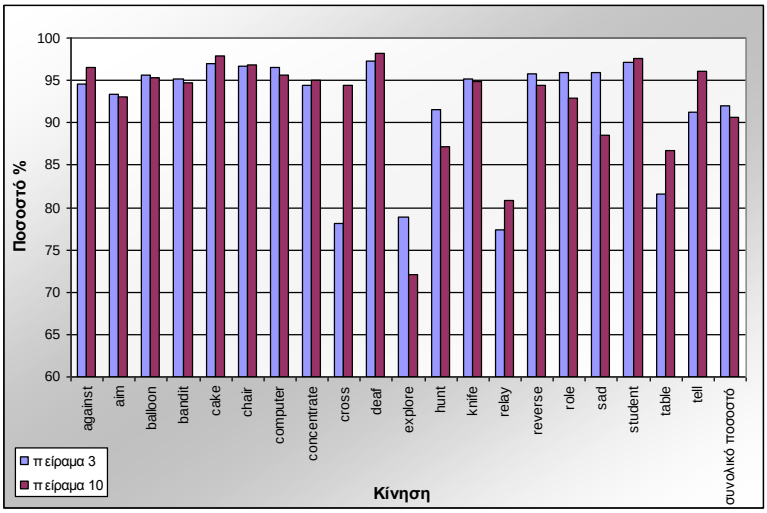
- Συχνά, όταν κάποιο από τα δύο χέρια είναι πλαγιασμένο ή με μικρό εμβαδόν, αναγνωρίζεται ως θόρυβος.
- Λόγω της ομοιότητας των δύο χεριών, η μεταξύ τους διαφοροποίηση (δεξί-αριστερό) έγκειται στην σχετική τους θέση στην εικόνα. Όταν όμως η κίνηση απαιτεί την εναλλαγή των δύο χεριών μεταξύ τους και αυτά διασταυρώνονται, το σύστημα αδυνατεί να τα αναγνωρίσει σωστά, μειώνοντας αρκετά το ποσοστό επιτυχίας στις κινήσεις αυτές.
- Στις κινήσεις, που προϋποθέτουν στροφή του κεφαλιού ή επικάλυψή του με ένα ή και τα δύο χέρια, το ποσοστό επιτυχίας μειώνεται, διότι το σύστημα αδυνατεί να αναγνωρίσει σωστά τις περιοχές αυτές.

Συνεχίζοντας, σε κάθε ένα από τα επόμενα πειράματα (5-8) χρησιμοποιήθηκαν δεδομένα εκπαίδευσης από βίντεο τριών ατόμων και ελέγχθηκε η δυνατότητα του αλγορίθμου να αναγνωρίσει τις περιοχές δέρματος από βίντεο του τέταρτου ατόμου. Για παράδειγμα τα δεδομένα εκπαίδευσης του πειράματος 5 προέρχονται από βίντεο των εκτελεστών 2-4. Τα αποτελέσματα των πειραμάτων αυτών απεικονίζονται στο σχήμα 5, όπου είναι σαφές ότι το ποσοστό αναγνώρισης κάποιου συγκεκριμένου εκτελεστή που έχει αποκλειστεί από τα δεδομένα εκπαίδευσης είναι μικρότερο, συγκριτικά με τους υπόλοιπους εκτελεστές που περιλαμβάνονται σε αυτά. Το προαναφερθέν όμως ποσοστό παραμένει υψηλό και κοντά στο μέσο όρο. Άρα, συμπεραίνουμε ότι μπορούμε να εκπαιδεύσουμε το σύστημα με δεδομένα από συγκεκριμένα πρόσωπα και αυτό να είναι ικανό να αναγνωρίσει σωστά περιοχές δέρματος από λήψεις βίντεο άλλων προσώπων.



Σχήμα 5: Ποσοστό % σωστά αναγνωρισμένων περιοχών ανά εκτελεστή των πειραμάτων 5-8

Τέλος, στο πείραμα 10 χρησιμοποιήσαμε σαν δεδομένα εκπαίδευσης το 60% των βίντεο των δέκα πρώτων κινήσεων και ελέγξαμε την αποδοτικότητα του αλγορίθμου στα υπόλοιπα δεδομένα. Τα αποτελέσματα του πειράματος αυτού απεικονίζονται ανά κίνηση ξεχωριστά στο σχήμα 6, σε σύγκριση με το πείραμα 3, όπου τα δεδομένα εκπαίδευσης είναι ίδιου μεγέθους, αλλά έχουνε εξαχθεί από βίντεο όλων των κινήσεων. Όπως είναι και φυσιολογικό, παρατηρούμε περιπτώσεις, όπου το ποσοστό των σωστά αναγνωρισμένων περιοχών στις τελευταίες δέκα κινήσεις που επιτυγχάνεται στο πείραμα 10 είναι μικρότερο από αυτό του πειράματος 3. Γενικά, όμως, μπορούμε να υποστηρίξουμε ότι το ποσοστό αυτό παραμένει σε όλες τις περιπτώσεις αρκετά υψηλό και συγκρίσιμο με το ποσοστό του πειράματος 3. Άρα, εν γένει, το σύστημα μπορεί να αναγνωρίσει σωστά περιοχές δέρματος, που προέρχονται από κινήσεις, που δεν έχουν χρησιμοποιηθεί στην κατασκευή των δεδομένων εκπαίδευσης.



Σχήμα 6: Ποσοστό % σωστά αναγνωρισμένων περιοχών ανά εκτελεστή των πειραμάτων 3 και 10

Ολοκληρώνοντας τα πειράματα ελέγχου του Bayesian δικτύου προκύπτουν κάποια γενικά συμπεράσματα ως προς την συμπεριφορά και την αποδοτικότητά του σε κάποιες περιπτώσεις κινήσεων, ατόμων και περιοχών. Αυτά συνοψίζονται στα εξής:

- Το δίκτυο είναι επιρρεπές σε λάθη αναγνώρισης σε περιπτώσεις στροφής του κεφαλιού, δηλαδή σε κινήσεις που η κινησιολογία τους απαιτεί το κεφάλι να μην είναι πάντα στραμμένο προς την κάμερα. Η στροφή του κεφαλιού επιφέρει μεγάλες διακυμάνσεις στις τιμές των ροπών με συνέπεια χαμηλότερα ποσοστά αναγνωρισιμότητας σε τέτοιες κινήσεις. Γενικά όμως, οι περισσότερες κινήσεις δεν απαιτούν στροφή του κεφαλιού.
- Το δίκτυο κάποιες φορές συγχέει την ταυτότητα των δύο χεριών όταν προηγουμένως στην κίνηση έχουν επικαλυφθεί μεταξύ τους ή με το κεφάλι. Επίσης, γενικότερα παρατηρείται ένα μειωμένο ποσοστό αναγνωρισιμότητας στις επικαλύψεις χεριών με το κεφάλι. Όταν η επικάλυψη σε μερικές κινήσεις είναι ολοκληρωτική, τότε οι τιμές των ροπών δεν διαφέρουν και άρα δεν επιτυγχάνεται αναγνώριση.
- Γενικά, σε κάθε πείραμα πέραν των μέσων όρων αναγνωρισιμότητας σε μία κίνηση ή σε ένα άτομο, υπάρχουν συγκεκριμένες λήψεις με σταθερά μικρότερη ή μεγαλύτερη απόδοση από τις υπόλοιπες του ίδιου ατόμου της ίδιας κίνησης. Άρα τα ποσοστά, ακόμα και τα υψηλά που εμφανίζονται, δεν είναι απόλυτα, αλλά περιέχουν κάποιες διακυμάνσεις μερικές φορές έντονες. Αυτό οφείλεται κυρίως σε εσφαλμένη σε κάποιο βαθμό ανίχνευση των περιοχών δέρματος, λόγω διαφορετικών συνθηκών φωτισμού και περιβάλλοντος γενικά, οι οποίες δεν είναι δυνατόν να διατηρηθούν ακριβώς ίδιες σε όλες τις λήψεις.

- Όσον αφορά την ικανότητα του δικτύου για αναγνώριση περιοχών άγνωστων κινήσεων, η απόδοσή του θεωρείται αρκετά καλή. Αποκλείοντας τις μισές κινήσεις από τα δεδομένα εκπαίδευσης, το ποσοστό αναγνώρισης των περιοχών τους ήταν φυσικά χαμηλότερο από αυτό των υπολοίπων περιοχών, κυμαινόμενο όμως σε υψηλά επίπεδα της τάξης του 85%.
- Αντίστοιχα, για την αναγνώριση περιοχών κινήσεων από άγνωστα άτομα τα ποσοστά επιτυχίας κυμάνθηκαν αρκετά υψηλά, τόσο ώστε να μπορούμε να ισχυριστούμε ότι το δίκτυο μπορεί κατά κανόνα να αναγνωρίσει περιοχές από κινήσεις άγνωστων ατόμων. Να σημειωθεί ότι σε κάθε πείραμα το άτομο, του οποίου οι λήψεις αποκλείονταν από την εκπαίδευση, εμφάνιζε μικρότερα ποσοστά επιτυχίας, όπως ήταν αναμενόμενο.
- Εν γένει, το δίκτυο εμφανίζει πολύ καλή συμπεριφορά στην περίπτωση μείωσης των δεδομένων εκπαίδευσης. Σε περιπτώσεις που αυτά είναι ισομερώς κατανεμημένα στις κινήσεις και στα άτομα τα αποτελέσματα που προκύπτουν για μικρό όγκο δεδομένων εκπαίδευσης διαφέρουν πολύ λίγο από αυτά για μεγάλο όγκο δεδομένων εκπαίδευσης.

2.4 Μελλοντική Έρευνα

Στην εργασία αυτή μελετήθηκε η απόδοση ενός Bayesian δικτύου στην αναγνώριση περιοχών και δείχθηκε μέσω πειραμάτων η υψηλή εν γένει αποτελεσματικότητά του στο πρόβλημα των επικαλύψεων, οι οποίες δύσκολα αντιμετωπίζονται από άλλες μεθόδους. Αυτό επιτεύχθηκε με την προσάρτηση ενός επιπλέον σταδίου επεξεργασίας.

Για περαιτέρω έρευνα στο πεδίο αυτό ακολουθώντας την μέθοδο των Bayesian δικτύων, παραθέτουμε κάποιες αλλαγές που ίσως αυξήσουν ακόμη περισσότερο την αξιοπιστία του. Τέτοιες είναι η επέκταση του δικτύου με επιπλέον κόμβους - μεταβλητές που θα περιγράφουν την πορεία συγκεκριμένης περιοχής κατά τη διάρκεια της κίνησης. Με την δημιουργία ενός μοντέλου κίνησης θα είναι εύκολη η πρόβλεψη της κίνησης σε μελλοντικά frame. Προσοχή όμως θα πρέπει να δοθεί στην αποτελεσματικότητα του δικτύου λαμβάνοντας υπόψη την υπολογιστική πολυπλοκότητα της λειτουργίας του.

Απώτερος στόχος είναι η δυνατότητα εξαγωγής μεταδεδομένων ακόμη μεγαλύτερου επιπέδου γνώσης, δηλαδή την αυτόματη αναγνώριση ολόκληρων κινήσεων νοηματικής γλώσσας και όχι μόνο περιοχών δέρματος. Μελλοντικά, μία μέθοδος σαν αυτή που παρουσιάστηκε θα ήταν χρήσιμη για τη υλοποίηση αυτού του στόχου.

3. Αναφορές

N. Ming-Hsuan Yang Ahuja, 1998, *Extracting gestural motion trajectories*, Dept. of Electrical & Computer Eng., Illinois Univ., Urbana, IL, Automatic Face and Gesture Recognition, Proceedings. Third IEEE International Conference.

P. K. Allen, Timcenko A., B. Yoshimi, P. Michelman, 1993, *Automated tracking and grasping of a moving object with a robotic hand-eye system*, IEEE Transactions on Robotics and Automation, 9(2):152-165.

F. Bashir, A. Khokhar, D. Schonfeld, 2005, *Automatic Object Trajectory-Based Motion Recognition using Gaussian Mixture Models*, University of Illinois, Chicago, IL.

R. Cutler and M. Turk, 1998, *View-based interpretation of real-time optical flow for gesture recognition*, Proc. Third IEEE Conference on Face and Gesture Recognition, Nara, Japan.

K. G. Derpanis, 2003, *Vision Based Gesture Recognition within a Linguistic Framework*, master's thesis, Master of Science, Graduate Program in Computer Science, York University, Toronto, Ontario .

S. Gong, Ng, J., Sherrah, J., 2002, *On the semantics of visual behavior, structured events and trajectories of human action*, Image and Vision Computing, 20, 873-888.

D. Heckerman, 1995, *A tutorial on learning with Bayesian networks*, Microsoft Research, MSR-TR-95-06.

C. Huang, A. Darwiche, 1996, *Inference in belief networks: A procedural guide*, International Journal of Approximate Reasoning, Volume15, Issue: 3, p. 225 - 263.

D. I. Kosmopoulos, I. Maglogiannis, 2006, *Extraction of Mid-Level Semantics from Gesture Videos using a Bayesian Network*, National Centre for Scientific Research "Demokritos", Institute of Informatics and Telecommunications, University of Aegian, Department of Information and Communications Systems Engineering.

Hyeon-Kyu Lee, Kim, J.H., 1999, *An HMM-Based Threshold Model Approach for Gesture Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10), 961.

K. Murphy, 1998, *A Brief Introduction to Graphical Models and Bayesian Networks*, http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html.

S. Park, J. K. Aggarwal, 2004, *Semantic-level Understanding of Human Action and Interactions using Event Hierarchy*, Electrical and Computer Engineering, University of Texas at Austin, USA.

J. Pearl, 1986, *Fusion, propagation and structuring in belief networks*, Artificial Intelligence, vol. 29, no. 3, pp. 241–288.

J. Rehg, Kanade, T., 1993, *DigitEyes: Vision-based Human Hand Tracking*, Tech. Rep. CMU-CS-93-220, School of Computer Science, Carnegie Mellon University, Pittsburg.

J. Shutler, 2002, *Complex Zernike Moments*, Statistical Moments, Department of Electronics and Computer Science, University of Southampton, United Kingdom, http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/SHUTLER3/node11.html.