УДК 621.391:004.9
doi:10.15217/issn1684-8853.2015.2.7

АНАЛИЗ СТРАТЕГИЙ И МЕТОДОВ ОБЪЕДИНЕНИЯ МНОГОМОДАЛЬНОЙ ИНФОРМАЦИИ

О. О. Басов[а], *канд. техн. наук, докторант*
А. А. Карпов[б], *доктор техн. наук, доцент*
[а]*Академия Федеральной службы охраны Российской Федерации, Орел, РФ*
[б]*Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, РФ*

Постановка проблемы: *в области искусственного интеллекта при проектировании многомодальных инфокоммуникационных систем и человеко-машинных интерфейсов крайне актуальны вопросы объединения разнородной информации (текстовой, акустической, визуальной и иных типов), поступающей как от пользователей, так и к ним по различным входным и выходным каналам коммуникации. Основная проблема при разработке и использовании многомодальных инфокоммуникационных систем заключается в том, что для них необходимы эффективные и надежные методы и технологии автоматического распознавания сигналов от каждой модальности, а также многомодального объединения информации и принятия решений.* ***Цель:*** *аналитический обзор научных методологических основ построения интеллектуальных инфокоммуникационных систем, опирающихся на многомодальные человеко-машинные интерфейсы.* ***Результаты:*** *представлен широкий спектр современной научно-технической литературы, описывающей результаты мировых научных исследований по данной теме за последнее десятилетие. Комплексный анализ существующих стратегий и математических методов обработки и интеграции многомодальной информации (на основе раннего, позднего и гибридного подходов к объединению), учета взаимной корреляции и синхронизации модальностей показал, что для большинства прикладных задач разработаны адекватные и эффективные способы объединения и разделения модальностей, которые должны грамотно применяться на этапе проектирования интеллектуальных систем.*

Ключевые слова *— инфокоммуникационная система, многомодальные интерфейсы, объединение и синхронизация модальностей, взаимная корреляция.*

Введение

Анализ структуры межличностной коммуникации и ее сопоставление с существующими и перспективными инфокоммуникационными системами позволяют выделить два общепризнанных подхода к представлению информации в таких системах. Первый из них, основанный на разделении передаваемой информации на услуги, часто применяется в существующих средствах телекоммуникации, однако не обеспечивает требуемой эффективности общения и, по большому счету, не имеет дальнейшего развития. Второй подход, реализующий многомодальное (полимодальное) представление информации, нашел широкое применение в информационных технологиях и, имея достаточно неплохие результаты от их применения, создал предпосылки к построению инфокоммуникационных систем на основе многомодальных пользовательских интерфейсов [1].

Основная проблема при их разработке связана с тем, что использование многомодального взаимодействия (*multimodal interaction*) человека с компьютером требует надежных методов распознавания сигналов от каждой модальности (канала передачи информации), а также эффективных способов объединения информации (*information fusion*) и принятия решений. Указанные обстоятельства обусловливают необходимость анализа стратегий и методов объединения входных и выходных потоков информации.

Стратегии объединения модальностей

Объединение многомодальной информации — это процесс, с помощью которого информация от различных информационных каналов интегрируется в единое представление. Объединение информации может происходить на различных уровнях ее представления и с использованием различных фундаментальных стратегий (рис. 1):

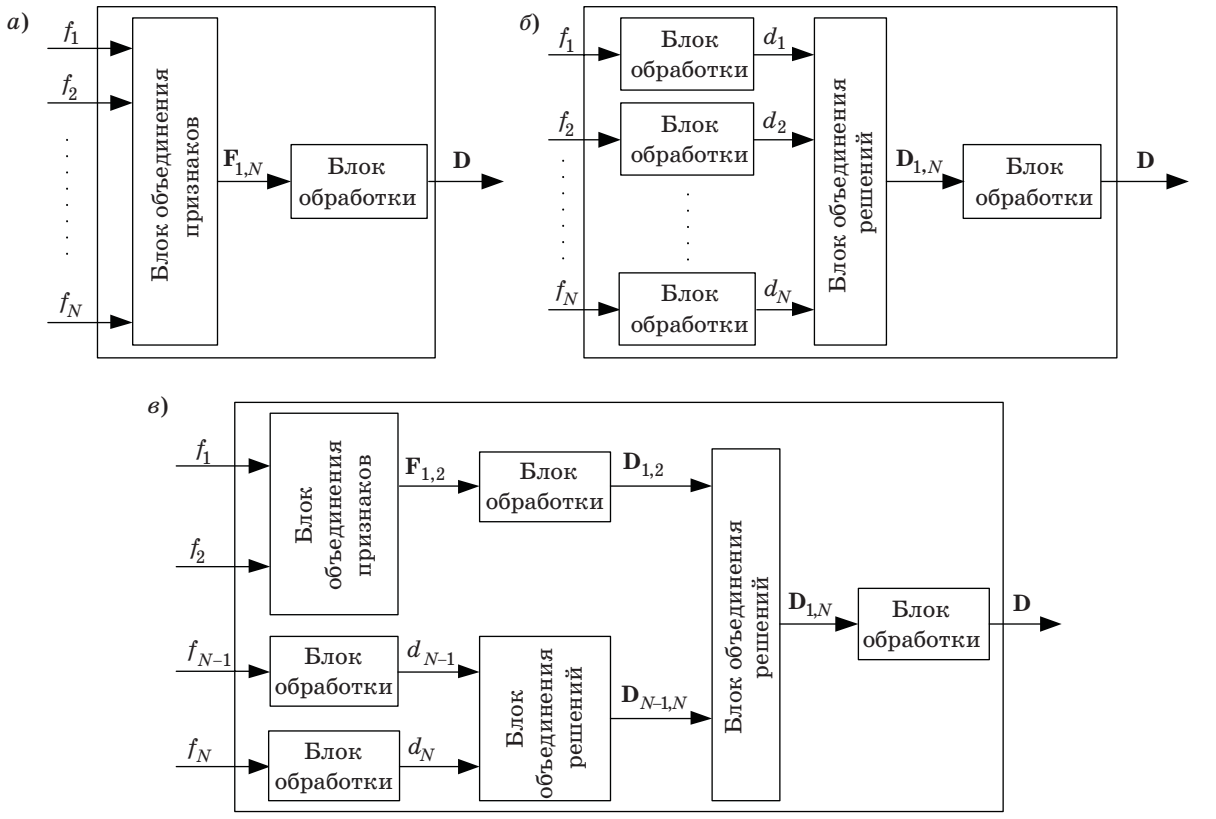
1) на уровне признакового описания (*feature level*), называемом «ранним объединением» (*early fusion*);

2) на (семантическом) уровне принятия решений (*decision level*), называемом «поздним объединением» (*late fusion*);

3) с использованием гибридного подхода (*hybrid approach*) [2].

В стратегии раннего объединения информативные признаки f_1, ..., f_N извлекаются из сигналов входных модальностей, интегрируются в векторе $\mathbf{F}_{1,N}$ и подаются в блок обработки, который формирует итоговое решение $\mathbf{D}$. Примером применения данной стратегии является задача поиска лица на изображении, в которой в качестве признаков используются цвет кожи и параметры характерных точек лица, а блок обработки представляет собой детектор лица.

При позднем объединении блоки обработки формируют локальные решения d_1, ..., d_N на основе соответствующих признаков f_1, ..., f_N. Локальные решения объединяются в вектор $\mathbf{D}_{1,N}$, на основе



■ ***Рис. 1.*** Стратегии объединения многомодальной информации: *а* — раннее; *б* — позднее; *в* — гибридное объединение

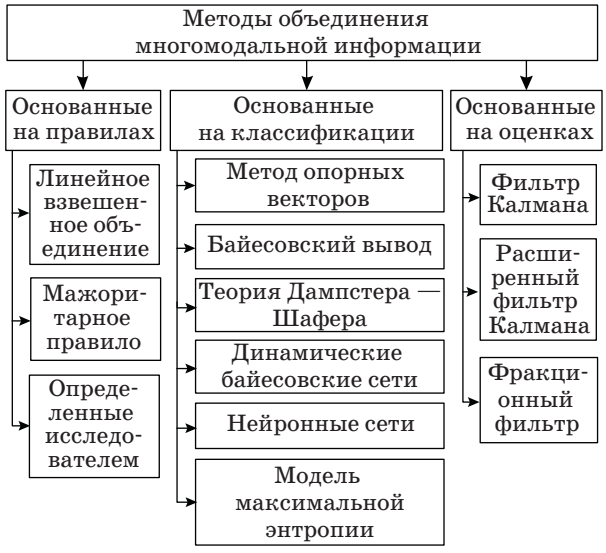
которого принимается итоговое решение **D** относительно решаемой задачи или выдвинутой гипотезы распознавания. Такой подход оправдывает себя при использовании методов анализа сигналов отдельных модальностей, например скрытых марковских моделей для аудиосигналов и метода опорных векторов для изображений, обеспечивая большую, чем при раннем объединении, гибкость обработки.

Стратегия гибридного объединения позволяет использовать достоинства обоих перечисленных выше подходов и также решать различные проблемы распознавания многомодальной информации.

Методы объединения многомодальной информации

Выбор стратегии объединения осуществляется в зависимости от требуемых функций многомодального человеко-машинного интерфейса, ограничений на способы ввода со стороны пользователя, а также предпочтительных способов объединения (рис. 2).

Методы многомодального объединения, основанные на правилах (*rule-based methods*), включают ряд основных норм комплексирования информации [3]. Линейное взвешенное объединение при ранней стратегии используется для обнаружения людей на изображении [4] и распознавания лиц [5], при стратегии позднего объединения — для распознавания диктора [6] и речи [7] с использованием акустического канала коммуникации, а также восстановления подвижных



■ ***Рис. 2.*** Классификация методов объединения модальностей

изображений по информации от текстового, акустического и визуального коммуникативных каналов [8]. Мажоритарное правило применено [9] для идентификации человека по информации от акустического и визуального каналов. Предложены методы позднего объединения текстового (ключевые слова) и визуального (цвет) коммуникативных каналов для индексирования спортивного видео [10], а также акустической (речь) и визуальной (2D- и 3D-жесты) информации для человеко-машинного взаимодействия [11].

Вторая группа методов объединения (см. рис. 2), основанных на классификации (*classification-based methods*), включает в себя ряд соответствующих технологий, используемых для соотнесения наблюдаемой многомодальной информации к предопределенным классам.

Метод опорных векторов (SVM) при позднем объединении информации текстового, акустического и визуального коммуникативных каналов применяется для определения семантики сообщения [12, 13], а текстового, визуального и тактильного — для биометрической верификации личности [14, 15]. Его использование при гибридном объединении информации акустического и визуального (цвет, размер, яркость, контраст) каналов коммуникации позволило системам анализировать мультимедиа [16] и классифицировать изображения [17], а также выполнять семантическую индексацию видео на основе текстовой и визуальной информации [18].

Байесовский вывод использован для распознавания речи при раннем [19] и позднем объединении [20] аудио и визуальных признаков, а также при гибридном объединении текстового, акустического и визуального каналов коммуникации для анализа спортивного видео [21]. Информация от акустического (коэффициенты линейного предсказания) и визуального (расположение и площадь объектов) каналов объединяется и для распознавания событий при наблюдении [22].

Применение теории Демпстера — Шафера при раннем [23] и гибридном [24] объединении визуальных признаков позволяет сегментировать изображения, а при позднем объединении информации акустического и визуального каналов коммуникации — классифицировать видео [25] и отпечатки пальцев [26], осуществлять человеко-машинное взаимодействие [27].

Динамические байесовские сети нашли широкое применение при раннем объединении акустических и визуальных признаков в задачах классификации кадров на видео [28], автоматического распознавания речи [29], локализации говорящего [30], сегментации новостного видео [31], биометрической верификации личности [32], локализации говорящего и слежения за объектами [33], классификации спортивного видео [34]. При позднем объединении визуальных признаков и параметров сенсоров (камер) динамические байесовские сети позволяют аннотировать фотографии [35] и отслеживать людей на изображениях [36]. Их применение при гибридном объединении информации текстового, акустического и визуального каналов коммуникации позволяет рубрицировать видео [37].

Раннее объединение информации акустического и визуального коммуникативных каналов с использованием искусственных нейронных сетей позволяет локализовать говорящего [38] и отслеживать людей на изображении [33]. Позднее объединение визуальных и тактильных признаков, а также параметров загрузки центрального процессора и сети на их основе позволяет осуществить мониторинг активности пользователя [39]. Гибридное объединение визуальной информации на основе искусственных нейронных сетей используется и при распознавании изображений [40].

Модель максимальной энтропии при раннем объединении информации текстового и визуального каналов коммуникации используется для индексации изображений [41].

Методы объединения информации третьей группы (см. рис. 2), основанные на количественных оценках (*estimation-based methods*), главным образом используются для отслеживания положения движущихся объектов (например, людей) на основе многомодальной информации. Так, например, фильтр Калмана и его модификации, а также фракционный фильтр на основе раннего и позднего объединения акустической и визуальной информации используются для определения положения [42, 43], отслеживания движений человека (объектов) [44, 45] и локализации дикторов [46, 47].

Анализ представленных методов многомодального объединения информации позволяет сделать следующие выводы:

— в настоящее время наибольшую распространенность среди методов объединения многомодальной информации получили метод опорных векторов и динамические байесовские сети;

— проанализированные методы чаще всего используются на уровне объединения векторов признаков (раннее объединение);

— самым вычислительно сложным методом являются динамические байесовские сети, наиболее простым — метод линейного взвешенного объединения;

— наиболее подходящими для решения проблем временных задержек и синхронизации модальностей различной природы являются специальные методы, основанные на пользовательских правилах.

Учет взаимной корреляции и синхронизация входных модальностей

Важную роль при объединении различных модальностей играет корреляция между ними, которая может определяться на различных уровнях объединения с использованием соответствующих методов. Так, в стратегии раннего объединения с применением:

— коэффициента корреляции решены задачи классификации видеокадров [48], распознавания речи [49], слежения за объектом [50], распознавания «говорящего» лица [51], стохастического кодирования видео и речевой информации [52];

— полного количества информации решены задачи распознавания речи [53], локализации говорящего [54] и слежения за диктором [55];

— латентного семантического анализа решены задачи распознавания «говорящего» лица [51] и биометрической аутентификации личности [56];

— канонического корреляционного анализа решены задачи биометрической аутентификации [56], распознавания «говорящего» лица [57] и верификации «говорящего» лица [58];

— кроссмодального факторного анализа осуществлен анализ мультимедийной виртуальной «говорящей головы» [59].

На базе стратегии позднего объединения решена задача распознавания событий при наблюдении с использованием коэффициента согласия [22] и анализа причинных связей [60].

При этом не только наличие корреляционных связей между отдельными модальностями, но и их независимость, в отдельных случаях, позволяет достичь лучших решений. Сигналы различных модальностей обычно фиксируются в различных форматах и с различной скоростью, в связи с чем возникает необходимость их синхронизации. При синхронизации на уровне признаков происходит объединение признаков, полученных в течение некоторого периода времени от разнородных, но сильно связанных и коррелированных по времени модальностей. Синхронизация на уровне принятия решения нуждается в определении некоторых временных меток, в которых решения будут объединяться. Таким образом, на обоих уровнях объединения модальностей проблема синхронизации возникает в различных формах.

Определение минимального времени синхронизации модальностей для различных приложений также остается актуальной задачей исследований в области построения многомодальных интерфейсов и систем, что подтверждается результатами и публикациями по данной проблематике [11, 15, 22, 30, 53, 61].

Заключение

Анализ разрабатываемых стратегий и методов объединения входных модальностей и результатов их применения позволяет утверждать, что большинство проводимых в мире исследований посвящено решению задач объединения модальностей, представленных в виде текстовой, аудио- (речь) и видеоинформации (2D и 3D). Применение правильных методов объединения информации при синтезе новых инфокоммуникационных систем на основе многомодальных интерфейсов обеспечит реализацию перцептивной стороны общения (познание друг друга партнерами по общению), а их дальнейшая интеллектуализация позволит приблизить инфокоммуникационное человеко-машинное взаимодействие к традиционному межличностному общению. Препятствиями этому служат нерешенные пока проблемы:

1) синхронизация отдельных модальностей в различных инфокоммуникационных приложениях;

2) динамическое определение оптимальных весовых коэффициентов различных модальностей;

3) моделирование и формализация процесса включения контекста при объединении многомодальной информации;

4) учет кроссмодальной корреляции при объединении информации на уровне принятия решений;

5) поиск более совершенных методов распознавания по каждой модальности;

6) оптимальный выбор методов объединения модальностей, учитывающий изменяющийся контекст, состояние и предпочтения пользователя, условия окружающей среды.

Для преодоления этих проблем требуется разработка строгой, но конструктивной теории, которая может стать научно-технологической основой для различных методов объединения и разделения многомодальной информации, позволяющих с единых методологических позиций оценивать существующее положение дел в предметной области и исследовать предлагаемые системотехнические решения по построению распределенных полимодальных инфокоммуникационных систем.

Данное исследование проводится при частичной финансовой поддержке Российского фонда фундаментальных исследований (проект № 15-07-04415) и Совета по грантам Президента РФ (грант № МД-3035.2015.8).

Литература

1. **Басов О. О., Саитов И. А.** Основные каналы межличностной коммуникации и их проекция на инфокоммуникационные системы // Тр. СПИИРАН. 2013. Вып. 30. С. 122–140.
2. **Atrey P. K., Hossain M. A., Kankanhalli M. S.** Multimodal Fusion for Multimedia Analysis: a Survey // Multimedia Systems. 2010. Vol. 16. Iss. 6. P. 345–379.
3. **Snoek C. G. M., Worring M., Smeulders A. W. M.** Early Versus Late Fusion in Semantic Video Analysis // Proc. ACM Intern. Conf. on Multimedia, Singapore, 2005. P. 399–402.
4. **Yang M. T., Wang S. C., Lin Y. Y.** A Multimodal Fusion System for People Detection and Tracking // Intern. Journal of Imaging Systems and Technology. 2005. N 15. P. 131–142.
5. **Kankanhalli M. S., Wang J., Jain R.** Experiential Sampling in Multimedia Systems // IEEE Transactions on Multimedia. 2006. N 5(8). P. 937–946.
6. **Neti C.,** et al. Joint Processing of Audio and Visual Information for Multimedia Indexing and Human-Computer Interaction // Proc. Intern. Conf. RIAO, France, 2000. P. 294–301.
7. **Карпов А. А.** Реализация автоматической системы многомодального распознавания речи по аудио- и видеоинформации // Автоматика и Телемеханика. 2014. Т. 75. № 12. С. 125–138.
8. **McDonald K., Smeaton A. F.** A Comparison of Score, Rank and Probability-Based Fusion Methods for Video Shot Retrieval // Proc. Intern. Conf. on Image and Video Retrieval, Singapore, 2005. P. 61–70.
9. **Radova V., Psutka J.** An Approach to Speaker Identification Using Multiple Classifiers // Proc. IEEE Intern. Conf. ICASSP, Munich, Germany, 1997. P. 1135–1138.
10. **Babaguchi N., Kawai Y., Kitahashi T.** Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration // IEEE Transactions on Multimedia. 2002. Vol. 4. P. 68–75.
11. **Holzapfel H., Nickel K., Stiefelhagen R.** Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3d Pointing Gestures // Proc. ACM Intern. Conf. on Multimodal Interfaces, USA, 2004. P. 175–182.
12. **Adams W.,** et al. Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues // EURASIP Journal on Applied Signal Processing. 2003. N 2. P. 170–185.
13. **Wu K., Lin C. K., Chang E., Smith J. R.** Multimodal Information Fusion for Video Concept Detection // Proc. IEEE Intern. Conf. on Image Processing, Singapore, 2004. P. 2391–2394.
14. **Aguilar J. F., Garcia J. O., Romero D. G., Rodriguez J. G.** A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification // Proc. Intern. Conf. on Video-Based Biometric Person Authentication AVBPA, Guildford, UK, 2003. P. 830–837.
15. **Bredin H., Chollet G.** Audio-visual Speech Synchrony Measure for Talking-Face Identity Verification // Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, Paris, France, 2007. Vol. 2. P. 233–236.
16. **Wu Y., Chang E. Y., Chang K. C. C., Smith J. R.** Optimal Multimodal Fusion for Multimedia Data Analysis // Proc. ACM Intern. Conf. on Multimedia, N. Y., USA, 2004. P. 572–579.
17. **Zhu Q., Yeh M. C., Cheng K. T.** Multimodal Fusion Using Learned Text Concepts for Image Categorization // Proc. ACM Intern. Conf. on Multimedia, S. Barbara, USA, 2006. P. 211–220.
18. **Ayache S., Quénot G., Gensel J.** Classifier Fusion for SVM-based Multimedia Semantic Indexing // Proc. 29th European Conf. on Information Retrieval Research, Rome, Italy, 2007. P. 494–504.
19. **Pitsikalis V., Katsamanis A., Papandreou G., Maragos P.** Adaptive Multimodal Fusion by Uncertainty Compensation // Proc. 9th Intern. Conf. Interspeech-2006, Pittsburgh, USA, 2006. P. 17–21.
20. **Meyer G., Mulligan J., Wuerger S.** Continuous Audio-visual Digit Recognition Using N-best Decision Fusion // Information Fusion. 2004. Vol. 5(2). P. 91–101.
21. **Xu H., Chua T. S.** Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video // ACM Transactions on Multimedia Computing Communications and Applications. 2006. Vol. 2(1). P. 44–67.
22. **Atrey P. K., Kankanhalli M. S., Jain R.** Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems // ACM/Springer Multimedia Systems Journal. 2006. Vol. 12(3). P. 239–253.
23. **Mena J. B., Malpica J.** Color Image Segmentation Using the Dempster–Shafer Theory of Evidence for the Fusion of Texture // Intern. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. 2003. Vol. XXXIV. Part 3/W8. P. 139–144.
24. **Bendjebbour A.,** et al. Multisensor Image Segmentation Using Dempster–Shafer Fusion in Markov Fields Context // IEEE Transactions on Geoscience and Remote Sensing. 2001. Vol. 39(8). P. 1789–1798.
25. **Guironnet M., Pellerin D., Rombaut M.** Video Classification Based on Low-Level Feature Fusion Model // Proc. 13th European Signal Processing Conf. EUSIPCO-2005, Antalya, Turkey, 2005. www.eurasip.org/Proceedings/Eusipco/Eusipco2005/defevent/papers/cr1344.pdf (дата обращения: 30.09.2014).
26. **Singh R., Vatsa M., Noore A., Singh S. K.** Dempster–Shafer Theory Based Finger Print Classifier Fusion with Update Rule to Minimize Training Time // IEICE Electronics Express. 2006. Vol. 3(20). P. 429–435.
27. **Reddy B. S.** Evidential Reasoning for Multimodal Fusion in Human Computer Interaction: Master of Science Thesis. — University of Waterloo, Canada. 2007. — 84 p.
28. **Wang Y., Liu Z., Huang J. C.** Multimedia Content Analysis: Using Both Audio and Visual Clues // IEEE Signal Processing Magazine. 2000. Vol. 17. Iss. 6. P. 12–36.

29. **Nefian A. V., Liang L., Pi X., Liu X., Murphye K.** Dynamic Bayesian Networks for Audio-visual Speech Recognition // EURASIP Journal on Advances in Signal Processing. 2002. N 11. P. 1–15.
30. **Nock H. J., Iyengar G., Neti C.** Speaker Localisation Using Audio-visual Synchrony: an Empirical Study // Proc. Intern. Conf. on Image and Video Retrieval, Urbana-Champaign, USA, 2003. P. 468–477.
31. **Chaisorn L.,** et al. A Multi-modal Approach to Story Segmentation for News // World Wide Web. 2003. N 6. P. 187–208.
32. **Hershey J., Attias H., Jojic N., Krisjianson T.** Audio Visual Graphical Models for Speech Processing // Proc. IEEE Intern. Conf. on Speech, Acoustics, and Signal Processing, Montreal, Canada, 2004. P. 649–652.
33. **Noulas A., Krose B.** EM Detection of Common Origin of Multi-modal Cues // Proc. Intern. Conf. on Multimodal Interfaces, Canada, 2006. P. 201–208.
34. **Ding Y., Fan G.** Segmental Hidden Markov Models for View-Based Sport Video Analysis // Proc. Intern. Workshop on Semantic Learning Applications in Multimedia, Minneapolis, USA, 2007. P. 1–8.
35. **Wu Y., Chang E., Tsengh B. L.** Multimodal Metadata Fusion Using Causal Strength // Proc. ACM Intern. Conf. on Multimedia, Singapore, 2005. P. 872–881.
36. **Town C.** Multi-sensory and Multi-modal Fusion for Sentient Computing // Intern. Journal of Computer Vision. 2007. Vol. 71. P. 235–253.
37. **Xie L.,** et al. Layered Dynamic Mixture Model for Pattern Discovery in Asynchronous Multi-modal Streams // Proc. IEEE Intern. Conf. ICASSP, USA, 2005. Vol. 2. P. 1053–1056.
38. **Cutler R., Davis L.** Look who's Talking: Speaker Detection Using Video and Audio Correlation // Proc. IEEE Intern. Conf. on Multimedia and Expo, New York City, USA, 2000. P. 1589–1592.
39. **Gandetto M.,** et al. From Multi-Sensor Surveillance Towards Smart Interactive Spaces // Proc. IEEE Intern. Conf. on Multimedia and Expo, USA, 2003. P. 641–644.
40. **Ni J., Ma X., Xu L., Wang J.** An Image Recognition Method Based on Multiple BP Neural Networks Fusion // Proc. IEEE Intern. Conf. on Information Acquisition. 2007. P. 429–435.
41. **Magalhaes J., Ruger S.** Information-Theoretic Semantic Multimedia Indexing // Proc. Intern. Conf. on Image and Video Retrieval, Amsterdam, 2007. P. 619–626.
42. **Andrieu C., Doucet A., Singh S., Tadic V.** Particle Methods for Change Detection, System Identification, and Control // Proc. of IEEE. 2004. Vol. 92(3). P. 423–438.
43. **Loh A., Guan F., Ge S. S.** Motion Estimation Using Audio and Video Fusion //Proc. Intern. Conf. on Control, Automation, Robotics and Vision. 2004. Vol. 3. P. 1569–1574.
44. **Gehrig T.,** et al. Kalman Filters for Audio-video Source Localization //Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Germany, 2005. P. 118–121.
45. **Talantzis F., Pnevmatikakis A., Polymenakos L. C.** Real Time Audio-visual Person Tracking // Proc. IEEE 8th Workshop on Multimedia Signal Processing, Victoria, USA, 2006. P. 243–247.
46. **Zotkin D. N., Duraiswami R., Davis L. S.** Joint Audio-visual Tracking Using Particle Filters // EURASIP Journal on Advances in Signal Processing. 2011. N 11. P. 1154–1164.
47. **Nickel K., Gehrig T., Stiefelhagen R., McDonough J.** A Joint Particle Filter for Audio-visual Speaker Tracking // Proc. 7th Intern. Conf. on Multimodal Interfaces, Trento, Italy, 2005. P. 61–68.
48. **Wang Y., Liu Z., Huang J. C.** Multimedia Content Analysis: Using both Audio and Visual Clues // IEEE Signal Processing Magazine. 2000. N 17(6). P. 12–36.
49. **Nefian A. V.,** et al. Dynamic Bayesian Networks for Audio-visual Speech Recognition // EURASIP Journal on Applied Signal Processing. 2002. N 11. P. 1–15.
50. **Beal M. J., Jojic N., Attias H.** A Graphical Model for Audiovisual Object Tracking// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003. N 25. P. 828–836.
51. **Li M., Li D., Dimitrove N., Sethi I. K.** Audio-visual Talking Face Detection // Proc. Intern. Conf. on Multimedia and Expo, Baltimore, USA, 2003. P. 473–476.
52. **Устинов А. А.** Стохастическое кодирование видео- и речевой информации: монография / под ред. В. Ф. Комаровича; Военная академия связи. — СПб., 2005. Ч. 1. — 220 с.
53. **Fisher-III J., Darrell T., Freeman W., Viola P.** Learning Joint Statistical Models for Audio-visual Fusion and Segregation // Advances in Neural Information Processing Systems. 2000. P. 772–778.
54. **Hershey J., Movellan J.** Audio-Vision: Using Audio-visual Synchrony to Locate Sounds // Advances in Neural Information Processing Syst. 2000. Vol. 12. P. 813–819.
55. **Iyengar G., Nock H. J., Neti C.** Audio-visual Synchrony for Detection of Monologue in Video Archives // Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, Hong Kong, 2003. P. I-329-32.
56. **Chetty G., Wagner M.** Audio-visual Multimodal Fusion for Biometric Person Authentication and Liveness Verification // Proc. NICTA-HCSNet Multimodal User Interaction Workshop, Sydney, Australia, 2006. P. 17–24.
57. **Slaney M., Covell M.** Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks // Proc. Neural Information Processing Society, Denver, USA, 2001. P. 814–820.
58. **Bredin H., Chollet G.** Audiovisual Speech Synchrony Measure: Application to Biometrics // EURASIP Journal on Advances in Signal Proc. 2007. P. 1–11.
59. **Li D., Dimitrova N., Li M., Sethi I. K.** Multimedia Content Processing Through Cross-Modal Association // Proc. ACM Intern. Conf. on Multimedia, Berkeley, USA, 2003. P. 2–5.

60. **Stauffer C.** Automated Audio-visual Activity Analysis// Technical report, MIT-CSAIL-TR-2005-057, USA, 2005. https://dspace.mit.edu/bitstream/handle/1721.1/30568/MIT-CSAIL-TR-2005-057.pdf?sequence=2 (дата обращения: 30.09.2014).

61. **Карпов А. А., Цирульник Л. И., Железны М.** Разработка компьютерной системы «говорящая голова» для аудиовизуального синтеза русской речи по тексту // Информационные технологии. 2010. Т. 9. № 8. С. 13–18.

UDC 621.391:004.9
doi:10.15217/issn1684-8853.2015.2.7

Analysis of Strategies and Methods for Multimodal Information Fusion

Basov O. O.[a], PhD., Tech., Doctoral Candidate, oobasov@mail.ru
Karpov A. A.[b], Dr. Sc., Tech., Associate Professor, karpov@iias.spb.su
[a]Academy of Federal Agency of Protection of Russian Federation, 35, Priborostroitelnaya St., 302034, Orel, Russian Federation
[b]Saint-Petersburg Institute of Informatics and Automation of RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation

Introduction: Design of multimodal infocommunication AI systems and human-computer interfaces often poses the problem of fusing various information (textual, acoustic, visual, etc.) which comes from or goes to the users via diverse input and output communication channels. The main problem in the development and application of multimodal infocommunication systems is providing efficient and reliable automatic recognition of signals of each modality, as well as multimodal fusion of information and decision making. **Purpose:** An analytical review of scientific methodological basis for the design of intellectual infocommunication systems based on multimodal human-computer interfaces. **Results:** A broad spectrum of modern research literature is presented, discussing the results in this domain published in the last decade all over the world. Complex analysis of state-of-the-art strategies and mathematical methods for multimodal information processing and integration (based on early, late and hybrid fusion strategies), taking into account the mutual correlation and synchronization of modalities, showed that for most actual applied problems we can find many adequate and efficient approaches for fusion and fission of modalities, which should be properly used at the stage of intelligent system design.
Keywords — Infocommunication System, Multimodal Interface, Fusion and Synchronization of Modalities, Mutual Correlation.

References

1. Basov O. O., Saitov I. A. Basic Channels of Interpersonal Communication and their Projection on the Infocommunications Systems. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2013, iss. 30, pp. 122–140 (In Russian).
2. Atrey P. K., Hossain M. A., Kankanhalli M. S. Multimodal Fusion for Multimedia Analysis: a Survey. *Multimedia Systems*, 2010, vol. 16, iss. 6, pp. 345–379.
3. Snoek C. G. M., Worring M., Smeulders A. W. M. Early Versus Late Fusion in Semantic Video Analysis. *Proc. ACM Intern. Conf. on Multimedia*, Singapore, 2005, pp. 399–402.
4. Yang M. T., Wang S. C., Lin Y. Y. A Multimodal Fusion System for People Detection and Tracking. *Intern. Journal of Imaging Systems and Technology*, 2005, no. 15, pp. 131–142.
5. Kankanhalli M. S., Wang J., Jain R. Experiential Sampling in Multimedia Systems. *IEEE Transactions on Multimedia*, 2006, no. 5(8), pp. 937–946.
6. Neti C., Maison B., Senior A., Iyengar G., Cuetos P., Basu S., Verma A. Joint Processing of Audio and Visual Information for Multimedia Indexing and Human-Computer Interaction. *Proc. Intern. Conf. RIAO, France*, 2000, pp. 294–301.
7. Karpov A. A. An Automatic Multimodal Speech Recognition System with Audio and Video Information. [Automation and Remote Control], 2014, vol. 75(12), pp. 2190–2200.
8. McDonald K., Smeaton A. F. A Comparison of Score, Rank and Probability-Based Fusion Methods for Video Shot Retrieval. *Proc. Intern. Conf. on Image and Video Retrieval*, Singapore, 2005, pp. 61–70.
9. Radova V., Psutka J. An Approach to Speaker Identification Using Multiple Classifiers. *Proc. IEEE Intern. Conf. ICASSP*, Munich, Germany, 1997, pp. 1135–1138.
10. Babaguchi N., Kawai Y., Kitahashi T. Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration. *IEEE Transactions on Multimedia*, 2002, vol. 4, pp. 68–75.
11. Holzapfel H., Nickel K., Stiefelhagen R. Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3d Pointing Gestures. *Proc. ACM Intern. Conf. on Multimodal Interfaces*, USA, 2004, pp. 175–182.
12. Adams W., Iyengar G., Lin C., Naphade M., Neti C., Nock H., Smith J. Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues. *EURASIP Journal on Applied Signal Processing*, 2003, no. 2, pp. 170–185.
13. Wu K., Lin C. K., Chang E., Smith J. R. Multimodal Information Fusion for Video Concept Detection. *Proc. IEEE Intern. Conf. on Image Processing*, Singapore, 2004, pp. 2391–2394.
14. Aguilar J. F., Garcia J. O., Romero D. G., Rodriguez J. G. A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification. *Proc. Intern. Conf. on Video-Based Biometric Person Authentication AVBPA*, Guildford, UK, 2003, pp. 830–837.
15. Bredin H., Chollet G. Audio-visual Speech Synchrony Measure for Talking-Face Identity Verification. *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, Paris, France, 2007, vol. 2, pp. 233–236.
16. Wu Y., Chang E. Y., Chang K. C. C., Smith J. R. Optimal Multimodal Fusion for Multimedia Data Analysis. *Proc. ACM Intern. Conf. on Multimedia*, New York, USA, 2004, pp. 572–579.
17. Zhu Q., Yeh M. C., Cheng K. T. Multimodal Fusion Using Learned Text Concepts for Image Categorization. *Proc. ACM Intern. Conf. on Multimedia*, S. Barbara, USA, 2006, pp. 211–220.
18. Ayache S., Quenot G., Gensel J. Classifier Fusion for SVM-based Multimedia Semantic Indexing. *Proc. 29th European Conf. on Information Retrieval Research*, Rome, Italy, 2007, pp. 494–504.
19. Pitsikalis V., Katsamanis A., Papandreou G., Maragos P. Adaptive Multimodal Fusion by Uncertainty Compensation. *Proc. 9th Intern. Conf. Interspeech-2006*, Pittsburgh, USA, 2006, pp. 17–21.
20. Meyer G., Mulligan J., Wuerger S. Continuous Audiovisual Digit Recognition Using N-best Decision Fusion. *Information Fusion*, 2004, vol. 5(2), pp. 91–101.
21. Xu H., Chua T. S. Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video. *ACM Transactions on Multimedia Computing Communications and Applications*, 2006, vol. 2(1), pp. 44–67.
22. Atrey P. K., Kankanhalli M. S., Jain R. Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems. *ACM/Springer Multimedia Systems Journal*, 2006, vol. 12(3), pp. 239–253.

23. Mena J. B., Malpica J. Color Image Segmentation Using the Dempster–Shafer Theory of Evidence for the Fusion of Texture. *Intern. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2003, vol. XXXIV, part 3/W8, pp. 139–144.
24. Bendjebbour A., Delignon Y., Fouque L., Samson V., Pieczynski W. Multisensor Image Segmentation Using Dempster–Shafer Fusion in Markov Fields Context. *IEEE Transactions on Geoscience and Remote Sensing*, 2001, no. 8(39), pp. 1789–1798.
25. Guironnet M., Pellerin D., Rombaut M. Video Classification Based on Low-level Feature Fusion Model. *Proc. 13th European Signal Processing Conf. EUSIPCO-2005*, Antalya, Turkey, 2005. Available at: www.eurasip.org/Proceedings/Eusipco/Eusipco2005/defevent/papers/cr1344.pdf (accessed 30 September 2014).
26. Singh R., Vatsa M., Noore A., Singh S. K. Dempster–Shafer Theory Based Finger Print Classifier Fusion with Update Rule to Minimize Training Time. *IEICE Electronics Express*, 2006, no. 3(20), pp. 429–435.
27. Reddy B. S. *Evidential Reasoning for Multimodal Fusion in Human Computer Interaction*. MS Thesis. University of Waterloo, Canada, 2007. 84 p.
28. Wang Y., Liu Z., Huang J. C. Multimedia Content Analysis: Using Both Audio and Visual Clues. *IEEE Signal Processing Magazine*, 2000, vol. 17, iss. 6, pp. 12–36.
29. Nefian A. V., Liang L., Pi X., Liu X., Murphye K. Dynamic Bayesian Networks for Audio-visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing*, 2002, no. 11, pp. 1–15.
30. Nock H. J., Iyengar G., Neti C. Speaker Localisation Using Audio-visual Synchrony: an Empirical Study. *Proc. Intern. Conf. on Image and Video Retrieval*, Urbana-Champaign, USA, 2003, pp. 468–477.
31. Chaisorn L., Chua T. S., Lee C. H., Zhao Y., Xu H., Feng H., Tian Q. A Multi-modal Approach to Story Segmentation for News. *World Wide Web*, 2003, no. 6, pp. 187–208.
32. Hershey J., Attias H., Jojic N., Krisjianson T. Audio Visual Graphical Models for Speech Processing. *Proc. IEEE Intern. Conf. on Speech, Acoustics, and Signal Processing*, Montreal, Canada, 2004, pp. 649–652.
33. Noulas A. K., Krose B. J. A. EM Detection of Common Origin of Multi-modal Cues. *Proc. Intern. Conf. on Multimodal Interfaces*, Banff, Canada, 2006, pp. 201–208.
34. Ding Y., Fan G. Segmental Hidden Markov Models for View-Based Sport Video Analysis. *Proc. Intern. Workshop on Semantic Learning Applications in Multimedia*, Minneapolis, USA, 2007, pp. 1–8.
35. Wu Y., Chang E., Tsengh B. Multimodal Metadata Fusion Using Causal Strength. *Proc. ACM Intern. Conf. on Multimedia*, Singapore, 2005, pp. 872–881.
36. Town C. Multi-Sensory and Multi-modal Fusion for Sentient Computing. *Intern. Journal of Computer Vision*, 2007, vol. 71, pp. 235–253.
37. Xie L., Kennedy L., Chang S. F., Divakaran A., Sun H., Lin C. Y. Layered Dynamic Mixture Model for Pattern Discovery in Asynchronous Multi-modal Streams. *Proc. IEEE Intern. Conf. ICASSP*, USA, 2005, vol. 2, pp. 1053–1056.
38. Cutler R., Davis L. Look who's Talking: Speaker Detection Using Video and Audio Correlation. *Proc. IEEE Intern. Conf. on Multimedia and Expo*, New York City, USA, 2000, pp. 1589–1592.
39. Gandetto M., Marchesotti L., Sciutto S., Negroni D., Regazzoni C. S. From Multi-sensor Surveillance Towards Smart Interactive Spaces. *Proc. IEEE Intern. Conf. on Multimedia and Expo*, USA, 2003, pp. 641–644.
40. Ni J., Ma X., Xu L., Wang J. An Image Recognition Method Based on Multiple BP Neural Networks Fusion. *Proc. IEEE Intern. Conf. on Information Acquisition*, 2007, pp. 429–435.
41. Magalhaes J., Ruger S. Information-Theoretic Semantic Multimedia Indexing. *Proc. Intern. Conf. on Image and Video Retrieval*, Amsterdam, 2007, pp. 619–626.
42. Andrieu C., Doucet A., Singh S., Tadic V. Particle Methods for Change Detection, System Identification, and Control. *Proc. of IEEE*, 2004, vol. 92(3), pp. 423–438.
43. Loh A., Guan F., Ge S. S. Motion Estimation Using Audio and Video Fusion. *Proc. Intern. Conf. on Control, Automation, Robotics and Vision*, 2004, vol. 3, pp. 1569–1574.
44. Gehrig T., Nicke K., Ekenel H., Klee U., McDonough J. Kalman Filters for Audio-video Source Localization. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Germany, 2005, pp. 118–121.
45. Talantzis F., Pnevmatikakis A., Polymenakos L. C. Real Time Audio-visual Person Tracking. *Proc. IEEE 8th Workshop on Multimedia Signal Processing*, Victoria, USA, 2006, pp. 243–247.
46. Zotkin D. N., Duraiswami R., Davis L. S. Joint Audio-visual Tracking Using Particle Filters. *EURASIP Journal on Advances in Signal Processing*, 2011, no. 11, pp. 1154–1164.
47. Nickel K., Gehrig T., Stiefelhagen R., McDonough J. Joint Audio-visual Tracking Using Particle Filters. *Proc. 7th Intern. Conf. on Multimodal Interfaces*, Trento, Italy, 2005, pp. 61–68.
48. Wang Y., Liu Z., Huang J. C. Multimedia Content Analysis: Using both Audio and Visual Clues. *IEEE Signal Processing Magazine*, 2000, vol. 17(6), pp. 12–36.
49. Nefian A. V., Liang L., Pi X., Liu X., Murphye K. Dynamic Bayesian Networks for Audio-visual Speech Recognition. *EURASIP Journal on Applied Signal Processing*, 2002, no. 11, pp. 1–15.
50. Beal M. J., Jojic N., Attias H. A Graphical Model for Audio-visual Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, no. 25, pp. 828–836.
51. Li M., Li D., Dimitrove N., Sethi I. K. Audio-visual Talking Face Detection. *Proc. Intern. Conf. on Multimedia and Expo*, Baltimore, USA, 2003. pp. 473–476.
52. Ustinov A. A. *Stokhasticheskoe kodirovanie video- i rechevoi informatsii* [Stochastic Coding of Video and Speech Information]. V. F. Komarovich ed. Vol. 1. Saint-Petersburg, VAS Publ., 2005. 220 p. (In Russian).
53. Fisher-III J., Darrell T., Freeman W., Viola P. Learning Joint Statistical Models for Audio-visual Fusion and Segregation. *Advances in Neural Information Processing Systems*, 2000, pp. 772–778.
54. Hershey J., Movellan J. Audio-vision: Using Audio-visual Synchrony to Locate Sounds. *Advances in Neural Information Processing Systems*, 2000, vol. 12, pp. 813–819.
55. Iyengar G., Nock H. J., Neti C. Audio-visual Synchrony for Detection of Monologue in Video Archives. *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003, pp. I-329-32.
56. Chetty G., Wagner M. Audio-visual Multimodal Fusion for Biometric Person Authentication and Liveness Verification. *Proc. NICTA-HCSNet Multimodal User Interaction Workshop*, Sydney, Australia, 2006, pp. 17–24.
57. Slaney M., Covell M. Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. *Proc. Neural Information Processing Society*, Denver, USA, 2001, pp. 814–820.
58. Bredin H., Chollet G. Audiovisual Speech Synchrony Measure: Application to Biometrics. *EURASIP Journal on Advances in Signal Processing*, 2007, pp. 1–11.
59. Li D., Dimitrova N., Li M., Sethi I. K. Multimedia Content Processing Through Cross-Modal Association. *Proc. ACM Intern. Conf. on Multimedia*, Berkeley, USA, 2003, pp. 2–5.
60. Stauffer C. *Automated Audio-visual Activity Analysis*. Tech. report, MIT-CSAIL-TR-2005-057, USA, 2005. Available at: www.eurasip.org/Proceedings/Eusipco/Eusipco2005/defevent/papers/cr1344.pdf (accessed 30 September 2014).
61. Karpov A. A., Tsirulnik L. I., Zelezny M. Development of a Computer System "Talking Head" for Text-to-Audiovisual-Speech Synthesis. *Informatsionnye tekhnologii* [Information Technologies], 2010, no. 8, pp. 13–18 (In Russian).