УДК 004.032.26

Э.А. Яндашевская

Разработка подсистемы стегоанализа цифровых изображений на основе сверточной нейронной сети для обнаружения и предовращения атак, использующих скрытые стеганографические каналы

Представлен вариант реализации подсистемы стегоанализа цифровых изображений, циркулирующих в информационной системе. Данная подсистема расширяет функциональность существующих систем обнаружения/предотвращения вторжений с точки зрения обнаружения скрытых каналов, применяемых в компьютерных атаках. В представленном варианте предложена и реализована параметрическая модель сверточной нейронной сети для обнаружения полезной нагрузки в цифровых изображениях, выполненных рядом распознанных в реальных атаках алгоритмах стеговложений. Разработана программная реализация модульного генератора обучающей выборки (датасета), поддерживающего эти алгоритмы. Осуществлена экспериментальная оценка точности.

Ключевые слова: защита информации, скрытые стеганографические каналы, сверточная нейронная сеть, системы обнаружения и предотвращения вторжений, цифровые изображения.
doi: 10.21293/1818-0442-2021-24-2-29-33

За период 2019/2020 годов компании, специализирующиеся на анализе инцидентов информационной безопасности, исследовали и провели технический анализ ряда специфических компьютерных атак, направленных, в первую очередь, на информационные инфраструктуры промышленных предприятий и коммерческих организаций. Целью этих атак является вымогательство либо путем блокирования доступа к информации, циркулирующей в атакуемых информационных системах, либо реализации техник перегрузки ресурсов вычислительных систем. Особенностью этих атак является использование на одном из этапов развития вектора атаки скрытых стеганографических каналов [1], базирующихся на цифровых изображениях (ЦИ), размещенных на легитимных публичных хостингах.

Так, в [2] проводится краткий технический анализ варианта такой атаки, направленной на десятки информационных систем предприятий и организаций в Японии, Италии, Германии и Великобритании с целью внедрения банковских троянов семейств Bebloh и Ursnif. В [3] приведен подробный технический анализ варианта этой же атаки, связанной с распространением трояна Ursnif. Схемой распространения вектора атаки, начинающейся рассылкой фишинговых электронных писем с вложением офисных документов, имеющих вредоносный VBA-скрип, является организация запросов к публичным хостингам ЦИ, таким как imgur.com и imgbox.com (рассмотрены в [2]) или postimg.cc (рассмотрен в [3]), с целью получения стегоконтейнеров, представленных легитимными ЦИ. В качестве стегоконтейнера используется ЦИ формата PNG. Стеговложение выполняется программой с открытым исходным кодом Invoke-PSImage [4], которая, получая на вход сценарий PowerShell, кодирует его байты в пиксели PNG-файла, используя модификацию метода LSB. В зависимости от варианта атаки такие ЦИ содержали либо обфурцированный скрипт PowerShell, реализующий дальнейшее развитие вектора атаки, либо непосредственно вредоносную полезную нагрузку. Этапы реализации указанных атак обобщены на рис. 1.

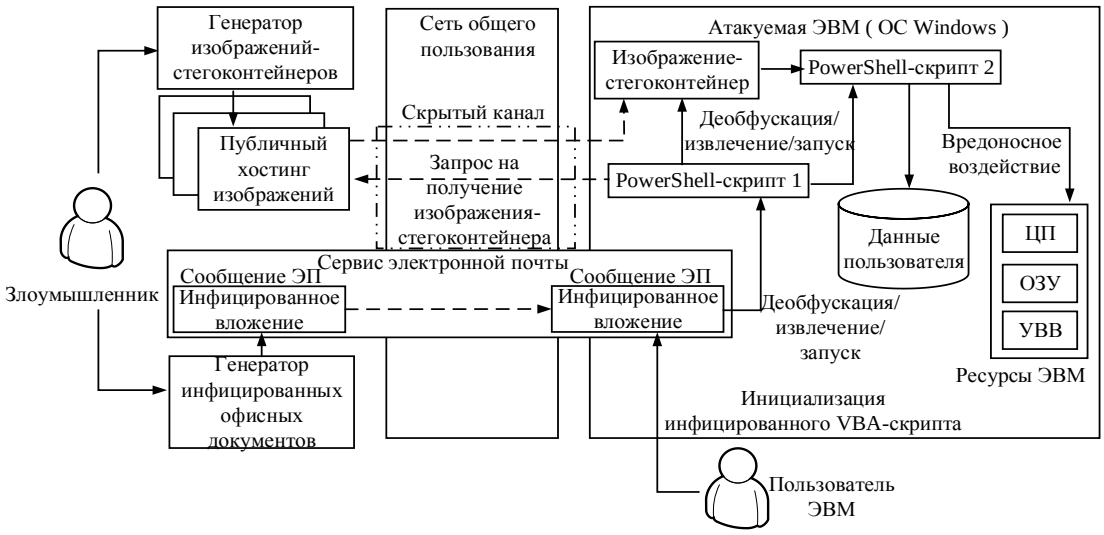

Рис. 1. Этапы компьютерных атак для распространения банковских троянов Bebloh и Ursnif со скрытым стеганографическим каналом на основе ЦИ формата PNG

Кроме того, в [5] приведен технический анализ атаки, выявленной летом 2020 г. Основой атаки является «MT3» (MontysThree) – многомодульный набор C++ инструментов, используемых для промышленного шпионажа. Как и в случае атак, представленных в [2, 3], их первым этапом является рассылка фишинговой корреспонденции с вредоносным вложением в офисные документы. На рис. 2 показана диаграмма взаимодействия модулей «MT3», из которой видно, что модуль ядра передается в виде полезной нагрузки в стегоконтенере ЦИ, представленном bitmap-файлом. Анализ указанного контейнера выявил многоэтапный процесс его распаковки и дешифрования. Особенностью модуля стеганографии в «MT3» является то, что его алгоритм является заказным, а не взят из стороннего репозитория с открытым исходным кодом.

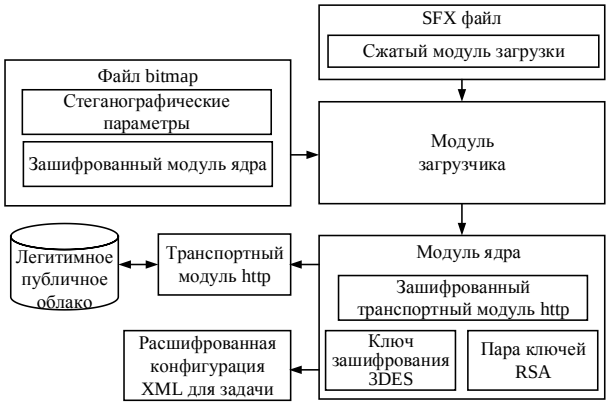

Рис. 2. Роль и место стеганографического контейнера, представленного bitmap-файлом, в наборе инструментов MontysThree [5]

Высокий процент успешных атак, выполненных на основе представленных примеров, показывает, что наряду с такими факторами, как социальная инженерия, их реализации способствовало сокрытие ряда компонентов в стегоконтейнерах ЦИ.

Поскольку системы защиты атакуемых предприятий и организаций, основанные на IDPS/DLP-сис-темах с традиционными для них контейнерным и сигнатурным видами анализа циркулирующей информации, не обнаруживают стегоконтейнеры ЦИ, относя канал их получения к легитимным каналам информационной системы, это делает задачу их обнаружения актуальной.

Решением этой проблемы может быть расширение функциональности IDPS/DLP-систем [6], включение в них модуля стегоанализа ЦИ с соответствующим перехватчиком, обеспечивающего решение задачи распознавания ЦИ – потенциальных стегоконтейнеров (рис. 3).

При этом важной исследовательской задачей является выбор метода распознавания, обеспечивающего оптимальную точность распознавания при низких временных и ресурсных издержках. Дополнительным условием является достаточная гибкость метода, позволяющая подстраиваться под модификации алгоритмов встраивания, используемых в реальных компьютерных атаках.

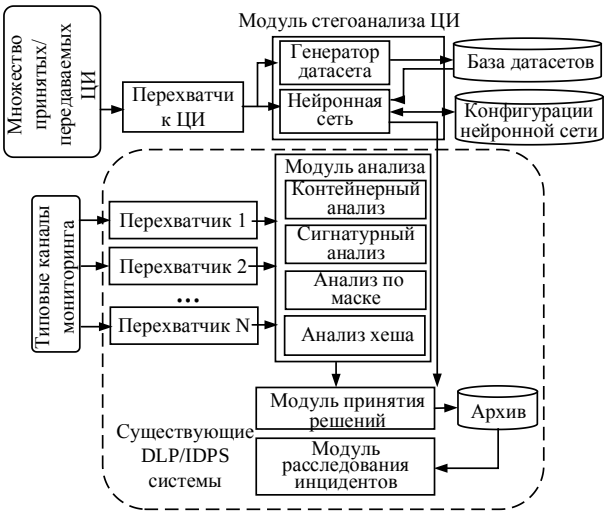

Рис. 3. Архитектура комбинированной IDPS/DLP-системы с дополнительным модулем стегоанализа ЦИ

Анализ существующих решений, используемых для обнаружения стеговложения

В настоящее время в зависимости от класса решаемой задачи распознавания стегоконтейнеров ЦИ существуют как исследовательские, так и коммерческие решения, основанные на двух видах классификаторов:

1. Двухэтапная классификация, основанная на моделях Rich Image Model (RIM) для пространственной или частотной областей ЦИ и статистических бинарных классификаторах, таких метод опорных векторов, линейный дискриминантный анализ или многослойный персептрон.

Примеры подобных проектов стегоанализаторов представлены в [7, 8].

Существенным недостатком таких решений является ограниченность используемых методов распознавания стеговложений к модификациям вариантов стегоконтейнеров. Частично, решение этой проблемы преодолевается использованием ансамбля классификаторов (Ensemble Classifier). Одно из таких решений представлено в [9].

2. Классификация на основе методов глубокого обучения, представленная глубокими сетями доверия (DBN), сверточными нейронными сетями (CNN), а также вариантами резервуарных вычислений, таких как эхо сети (ECN) и машины жидкостного состояния (LSM). Реализации проектов на основе указанных методов представлены в [10]. Достоинством методов глубокого обучения является возможность их переобучения на распознавание конкретного набора стегоконтейнеров ЦИ с сохранением структурно-параметрических характеристик нейронной сети для повторного использования. К недостаткам стоит отнести достаточно сложный этап обучения сети [11], требующий тщательного подбора обучающей выборки ЦИ – датасета.

Вариант модуля стегоанализа цифровых изображений на основе сверточной нейронной сети

В исследовании в качестве основы модуля стегоанализа ЦИ предлагается использование варианта

сверточной нейронной сети (СНС). Обоснование такого выбора и достоинства ее применения представлены в работе автора [12]. Структура разработанной СНС в общем случае соответствует представленной в [13] архитектуре сверточных сетей и представлена следующими типами слоев: сверточные слои; слои подвыборки (pooling); полносвязные слои.

Сверточные слои, начиная с первого, распознают низкоуровневые признаки ЦИ. По мере продвижения по ним эти признаки обобщаются, что позволяет переходить к высокоуровневым признакам ЦИ. Базовая функция сверточного слоя и слоя подвыборки уменьшение ядра – матрицы весов. Их параметрами являются: f (filters count) – количество фильтров в слое; K (kernel size) – размер (высота и ширина) ядра; s (stride) – шаг свёртки (количество пикселей, на которое перемещается матрица фильтра по входному изображению); p (padding) – дополнения нулями (количество пикселей, которые добавляются с каждого края изображения). Перемещением ядра над пикселями ЦИ выполняется перемножение и последующее суммирование его весов и значение пикселей, над которыми находится ядро. Эта функция именуется двумерной сверткой. Ее результатом для последующего сверточного слоя новое ядро меньшего размера. Сверточные слои и слои подвыборки используются для предварительной обработки ЦИ. В таблице представлены значения параметров сверточных слоев и слоев подвыборки разработанной СНС.

Значения параметров сверточных слоев и слоев подвыборки разработанной СНС

Слой	Значение f (in/out)	Размер K	Значение s	Значение p
Сверточный слой 1	3/8	5	1	2
Сверточный слой 2	8/16	5	1	2
Сверточный слой 3	16/64	5	1	2
Слой подвыборки 1		5	4	2
Слой подвыборки 2		5	4	2

Полносвязный слой реализует вариант нелинейной функции, которая, проверяя комбинации входных данных, реализует бинарную (Cover/Stego) классификацию ЦИ. В разработанной СНС реализовано 5 полносвязных слоев, редуцирующих значения f (in/out) с 4096/2048 до 128/2, где 2 – результат решения задачи бинарной классификации.

Важной задачей при разработке СНС явилась задача выбора обучающей выборки – датасета. Ее особенностью применительно к решаемой задаче явилось то, что большинство исследований в области разработки СНС-стегоанализоаторов ЦИ в качестве базы ЦИ используют известные проекты, такие, например, как BOSS (Break Our Steganographic System) [14] и RAISE (The Raw Images Dataset) [15], применяя для формирования стегоконтейнеров известные варианты алгоритмов с высоким соотношением скрытности к объему полезной нагрузки, такие как steGO (HUGO), WOW, UNIWARD, STABYLO, EAI-LSBM, MVG [16].

Однако, как было рассмотрено выше, в реальных реализациях атак с использованием стегоконтейнеров для последних применяются специфические для вида атаки алгоритмы стеговложения. Примером является рассмотренный выше алгоритм Invoke-PSImage, реализующий стеговложение строго для сценариев PowerShell. В связи с этим было принято решение реализовать генератор датасета СНС, имеющий модульную структуру и позволяющий подключать распознанные алгоритмы стеговложений реальных компьютерных атак. В качестве базы ЦИ было предложено использовать наиболее известные публичные хостинги ЦИ, также применяемые в ходе этих атак.

Результаты эксперимента

Обучение СНС выполнялось на массиве из 35000 ЦИ, полученных из публичного хостинга imgur.com. На базе 15000 из них, с использованием алгоритма Invoke-PSImage, были реализованы PNG-стегоконтейнеры с различной полезной нагрузкой. Тестирование обученной СНС производилось на 2000 ЦИ, из которых 1000 были представлены вариантами Invoke-PSImage стегоконтейнеров. И в процессе обучения, и в процессе тестирования размер батча (подмножества ЦИ, подаваемого на вход СНС) был равен 64. На рис. 4 представлены графики точности распознавания (accuracy) и функции потерь для этапов обучения и тестирования СНС.

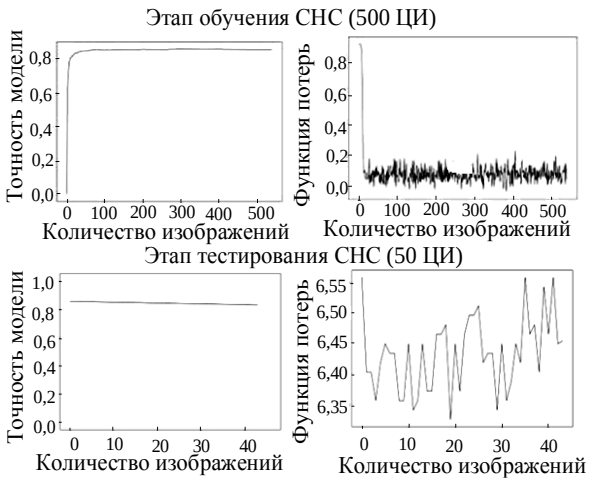

Рис. 4. Результаты экспериментальной оценки разработанной СНС

Заключение

В статье на примерах зафиксированных компьютерных атак с использованием скрытых стеганографических каналов предложен подход по расширению функциональности современных систем обнаружения и предотвращения вторжений за счет использования модуля стегоанализа цифровых изображений, циркулирующих в защищаемой ин-

формационной системе. В качестве базы такого модуля обоснован выбор аппарата сверточных нейронных сетей. Предложен вариант сверточной нейронной сети, а также модульный генератор датасета для нее, обеспечивающий подключение распознанных в реальных атаках алгоритмов стеговложений. На основе сформированных обучающей и тестовой выборок цифровых изображений, полученных с использованием в компьютерных атаках публичного хостинга, проведен эксперимент, демонстрирующий приемлемую степень распознавания. Направлением дальнейших исследований является совершенствование предложенного стегоанализатора путем комбинирования методов статистической бинарной классификации и сверточных нейронных сетей с целью сокращения временных и ресурсных затрат, связанных с этапом обучения сети.

Литература

1. Яндашевская Э.А. Подход к тестированию на проникновение в информационные сервисы сети RSNet по скрытым каналам, основанным на методах стеганографического преобразования информации // Науч.-техн. журнал «I-Methods» / ООО «Институт ИНТЕХ». – 2020. – Т. 12, № 4. – С. 1–14.

2. Стеганография в атаках на промышленные предприятия (обновлено) // Лаборатория Касперского. – 2020 [Электронный ресурс]. – Режим доступа: https://icscert.kaspersky.ru/reports/2020/06/17/steganography-in-attacks-on-industrial-enterprises/, свободный (дата обращения: 21.01.2021).

3. Paqanin P. Ursnif: Long Live the Steganography and AtomBombing! // Security affairs. – 2019 [Электронный ресурс]. – Режим доступа: https://securityaffairs.co/wordpress/80777/malware/ursnif-steganography-atombombing.html (дата обращения: 21.01.2021).

4. InvokePSImage source code [Электронный ресурс]. – Режим доступа: https://github.com/peewpw/Invoke-PSImage, свободный (дата обращения: 15.10.2020).

5. Legezo D. MontysThree: Industrial espionage with steganography and a Russian accent on both sides. – 2020 [Электронный ресурс]. – Режим доступа: https://securelist.com/montysthree-industrial-espionage/98972/, свободный (дата обращения: 15.10.2020).

6. Белова А.Л. Сравнительный анализ систем обнаружения вторжений / А.Л. Белова, Д.А. Бородавкин // Актуальные проблемы авиации и космонавтики. – 2016. – № 1. – С. 742–744.

7. Fridrich J.J. Rich models for steganalysis of digital images / J.J. Fridrich, J. Kodovsk´y // IEEE Trans Inform Forensics Secur. – 2012. – Vol. 7, No. 3. – P. 868–882.

8. Holub V. Random projections of residuals for digital image steganalysis / V. Holub, J.J. Fridrich // IEEE TransInform Forensics Security. – 2013. – Vol. 8, No. 12. – P. 1996–2006.

9. Kodovsk´y J. Ensemble classifiers for steganalysis of digital media / J. Kodovsk´y, J.J. Fridrich, V. Holub // IEEE Trans Inform Forensics Security – 2012. – Vol. 7, No. 2. – P. 432–444.

10. Maass W. Liquid state machines: motivation, theory, and applications // Computability in context: computation and logic in the real world. – 2010. – С. 275–296.

11. Moiseev O.V. To the question of sensitivity checking of the conventional neural network for digital images stegoanalysis to the sampling parameters at the learning stage / O.V. Moiseev, E.A. Yandashevskaya // MIP2021'SCT: Proceedings of the XXVI-th International Open Science. Yelm, WA, USA: Science Book Publishing House. – 2021. – P. 59–63.

12. Полунин А.А. Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений / А.А. Полунин, Э.А. Яндашевская // Труды ИСП РАН. – 2020. – № 32(4). – С. 155–164.

13. Qian Y. Deep learning for steganalysis via convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // IS&T/SPIE Electronic Imaging. – 2015. DOI: 10.1117/12.2083479.

14. Be the boss of the BOSS, Break Our Steganographic System! BOSS – Break Our Steganographic System [Электронный ресурс]. – Режим доступа: http://agents.fel.cvut.cz/boss/index.php?mode=VIEW&tmpl=materials, свободный (дата обращения: 10.11.2020).

15. RAISE – RAW Images Dataset [Электронный ресурс]. – Режим доступа: http://loki.disi.unitn.it/RAISE/, свободный (дата обращения: 10.11.2020).

16. Pevn´y T. Using high-dimensional image models to perform highly undetectable steganography / T. Pevn´y, T. Filler, P. Bas // Information Hiding – 12th International Conference. – 2010. – P. 161–177.

Яндашевская Элина Андреевна
Сотрудник ФГКВОУВО «Академия Федеральной службы охраны Российской Федерации»
Приборостроительная ул., 35, г. Орел, Россия, 302025
Тел.: +7 (486-2) 54-99-33
Эл. почта: elenayanda@yandex.ru

Yandashevskaya E.A.

Development of a Subsystem for Steganalysis of Digital Images Based on a Convolutional Neural Network to Detect and Prevent Attacks Using Hidden Steganographic Channels

This article presents a way to implement the subsystem for steganalysis of digital images circulating in the information system. This subsystem expands the functionality of existing intrusion detection / prevention systems in terms of detecting covert channels used in computer attacks. In the presented solution, a parametric model of a convolutional neural network is proposed and implemented to detect a payload in digital images, performed by a number of steg-nesting algorithms recognized in real attacks. A software implementation of a modular generator of a training sample (dataset) that supports these algorithms has been developed. An experimental assessment of the accuracy has been carried out.

Keywords: information protection, hidden steganographic channels, convolutional neural network, intrusion detection and prevention systems, digital images.
doi: 10.21293/1818-0442-2021-24-2-29-33

References

1. Yandashevskaya E.A. [An approach to penetration testing of information services of the RSNet network through covert channels based on methods of steganographic transformation of information]. *Scientific and Technical Journal «I-Methods». Institute INTECH LLC*, 2020, vol. 12, no. 4, pp. 1–14 (in Russ.).

2. Steganography in attacks on industrial enterprises (updated). *Kaspersky Lab*, 2020. (In Russ.). Available at: https://icscert.kaspersky.ru/reports/2020/06/17/steganography-

in-attacks-on-industrial-enterprises/, free (Accessed: January 21, 2021).

3. Paqanin P. Ursnif: Long Live the Steganography and AtomBombing! *Security Affairs*, 2019. Available at: https://securityaffairs.co/wordpress/80777/malware/ursnif-steganography-atombombing.html, free (Accessed: January 21, 2021).

4. InvokePSImage source code. *GitHub*, 2019. Available at: https://github.com/peewpw/Invoke-PSImage, free (Accessed: October 15, 2020).

5. Legezo D. MontysThree: Industrial espionage with steganography and a Russian accent on both sides. *Securelist by Kaspersky*, 2020. Available at: https://securelist.com/montysthree-industrial-espionage/98972/, free (Accessed: October 15, 2020).

6. Belova A.L., Wartkin D.A. [Comparative analysis of intrusion detection systems]. *Actual Problems of Aviation and Astronautics*, 2016, no. 1, pp. 742–744 (in Russ.).

7. Fridrich J.J., Kodovsk´y J. [Rich models for steganalysis of digital images]. *IEEE Transactions on Information Forensics and Security*, 2012, vol. 7, no. 3, pp. 868–882.

8. Holub V., Fridrich J.J. [Random projections of residuals for digital image steganalysis]. *IEEE Transactions on Information Forensics and Security,* 2013, no. 12, pp. 1996–2006.

9. Kodovsk´y J., Fridrich J.J., Holub V. [Ensemble classifiers for steganalysis of digital media]. *IEEE Transactions on Information Forensics and Security,* 2012, vol. 7, no. 2, pp. 432–444.

10. Maass W. [Liquid state machines: motivation, theory, and applications]. *Computability in Context: Computation and Logic in the Real World*, 2010, pp. 275–296.

11. Moiseev O.V., Yandashevskaya E.A. [To the question of sensitivity checking of the conventional neural network for digital images stegoanalysis to the sampling parameters at the learning stage] *MIP2021'SCT: Proceedings of the XXVI-th International Open Science Conference*. Editor in Chief Dr. Sci., Prof. O.Ja. Kravets. Yelm, WA, USA: Science Book Publishing House, 2021, pp. 59–63 (in Russ.).

12. Polunin A.A., Yandashevskaya E.A. [Using the apparatus of convolutional neural networks for steganalysis of digital images]. *Proceedings of ISP RAS*, 2020, vol. 32, no. 4, pp. 155–164 (in Russ.).

13. Qian Y., Dong J., Wang W., Tan T. [Deep learning for steganalysis via convolutional neural networks]. *IS&T/SPIE Electronic Imaging*, 2015. doi: 10.1117/12.2083479

14. Be the boss of the BOSS, Break Our Steganographic System! Available at: http://agents.fel.cvut.cz/boss/index.php?mode=VIEW&tmpl=materials, free (Accessed: November 10, 2020).

15. RAISE – RAW Images Dataset. Introducing RAISE dataset. Available at: http://loki.disi.unitn.it/RAISE/, free (Accessed: November 10, 2020).

16. Pevn´y T., Filler T., Bas P. [Using high-dimensional image models to perform highly undetectable steganography]. *Information Hiding – 12th International Conference*, 2010, pp. 161–177.

Elina A. Yandashevskaya
Employee, Academy of the Federal Guard Service
35, Priborostroitelnaya st., Orel, Russia, 302034
Phone: +7 (486-2) 54-99-33
Email: elenayanda@yandex.ru